ENCODE

4D Nucleome

3D Genome Organization workshop

Feng Yue, PhD
Co-Chair, ENCODE Outreach Group
Member, 4DN Outreach Group
Dept. of Biochemistry & Molecular Biology
College of Medicine
Pennsylvania State University
Vancouver, ASHG 2016

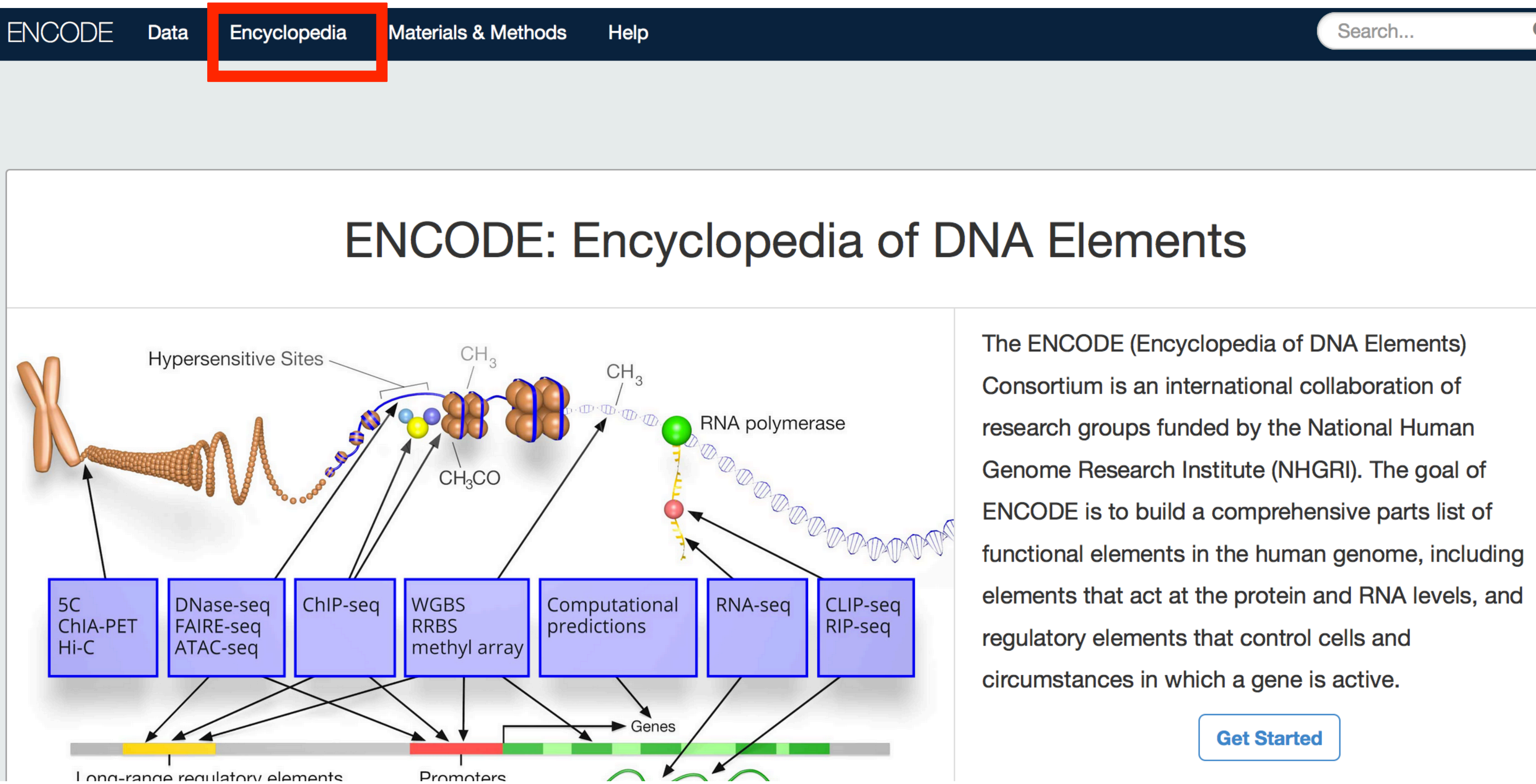
Goal of this workshop

- Tell the community about the resources that are/ will be available through ENCODE and 4DN projects;
- Hear feedbacks from users what we can do to make our data more usable and friendly;
- Slides will be available through ENCODE portal and 4DN websites.

Outline

- Dan Gilchrist (NHGRI, ENCODE)
- Lisa Chadwick (NIEHS, Roadmap/Epigenomics, 4DN)
- David Gorkin (UCSD, Bing Ren’s Group)
- Feng Yue (Penn State, ENCODE, 4DN)
- Ting Wang (Wash. U., Roadmap, 4DN)

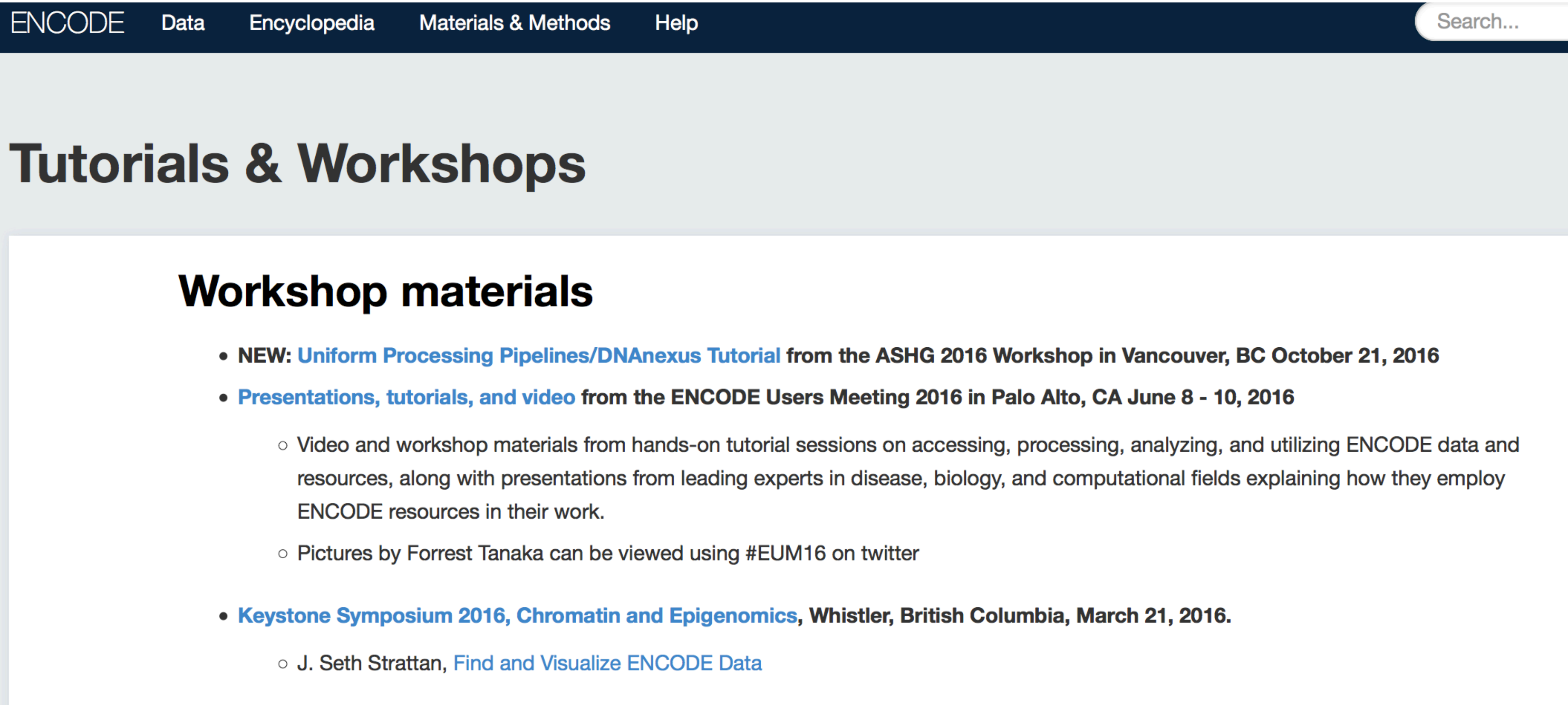
encodeproject.org



ENCODE: Encyclopedia of DNA Elements

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Tutorials available on the ENCODE portal

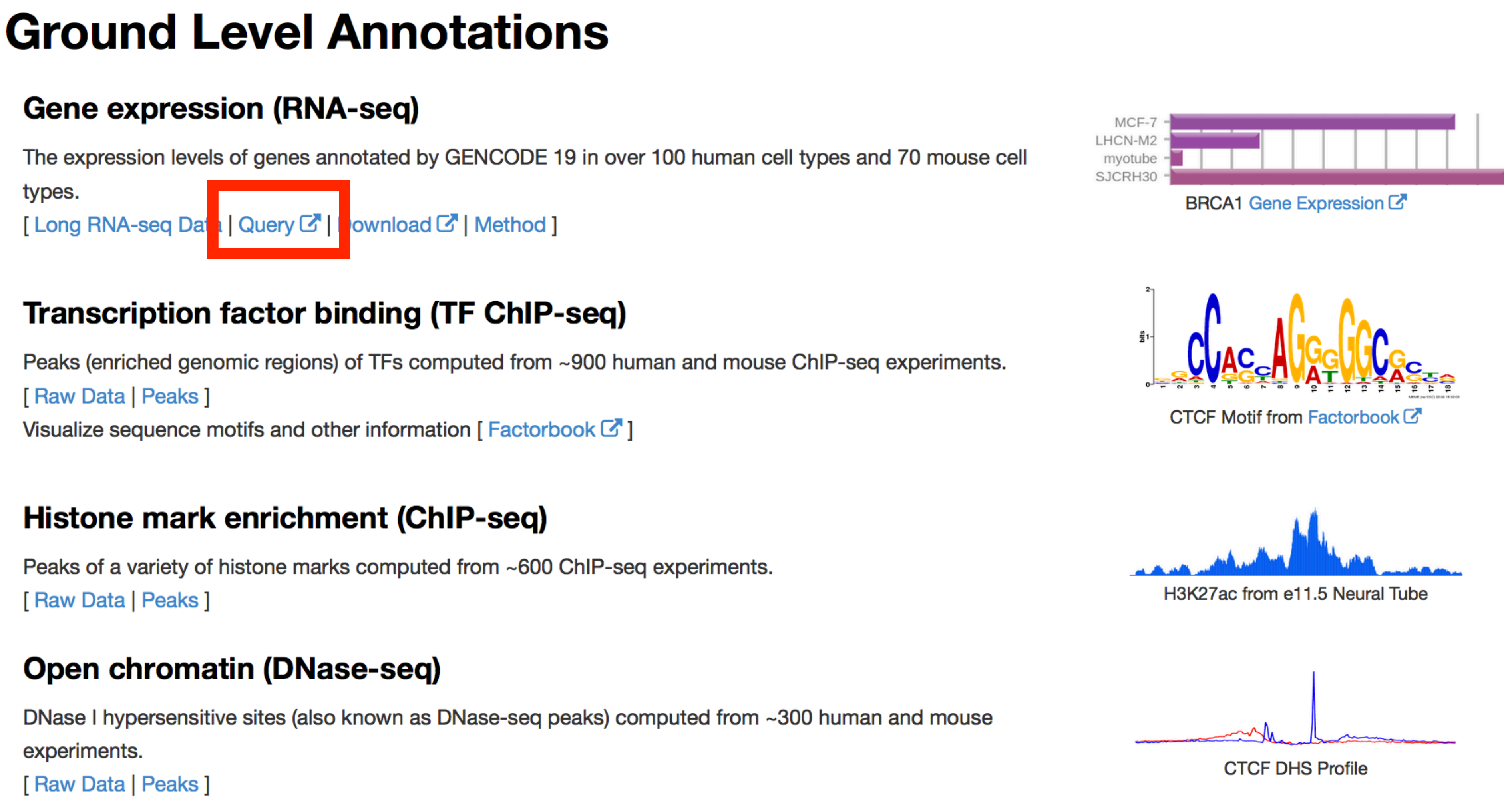


Tutorials & Workshops

Workshop materials

- **NEW: Uniform Processing Pipelines/DNAnexus Tutorial from the ASHG 2016 Workshop in Vancouver, BC October 21, 2016**
- **Presentations, tutorials, and video from the ENCODE Users Meeting 2016 in Palo Alto, CA June 8 - 10, 2016**
  - Video and workshop materials from hands-on tutorial sessions on accessing, processing, analyzing, and utilizing ENCODE data and resources, along with presentations from leading experts in disease, biology, and computational fields explaining how they employ ENCODE resources in their work.
  - Pictures by Forrest Tanaka can be viewed using #EUM16 on twitter
- **Keystone Symposium 2016, Chromatin and Epigenomics, Whistler, British Columbia, March 21, 2016.**
  - J. Seth Strattan, Find and Visualize ENCODE Data

ENCODE element query website



Ground Level Annotations

Gene expression (RNA-seq)

The expression levels of genes annotated by GENCODE 19 in over 100 human cell types and 70 mouse cell types.

[Long RNA-seq Data | Query | Download | Method]

BRCA1 Gene Expression

Transcription factor binding (TF ChIP-seq)

Peaks (enriched genomic regions) of TFs computed from ~900 human and mouse ChIP-seq experiments.

[Raw Data | Peaks]

Visualize sequence motifs and other information [Factorbook]

CTCF Motif from Factorbook

Histone mark enrichment (ChIP-seq)

Peaks of a variety of histone marks computed from ~600 ChIP-seq experiments.

[Raw Data | Peaks]

H3K27ac from e11.5 Neural Tube

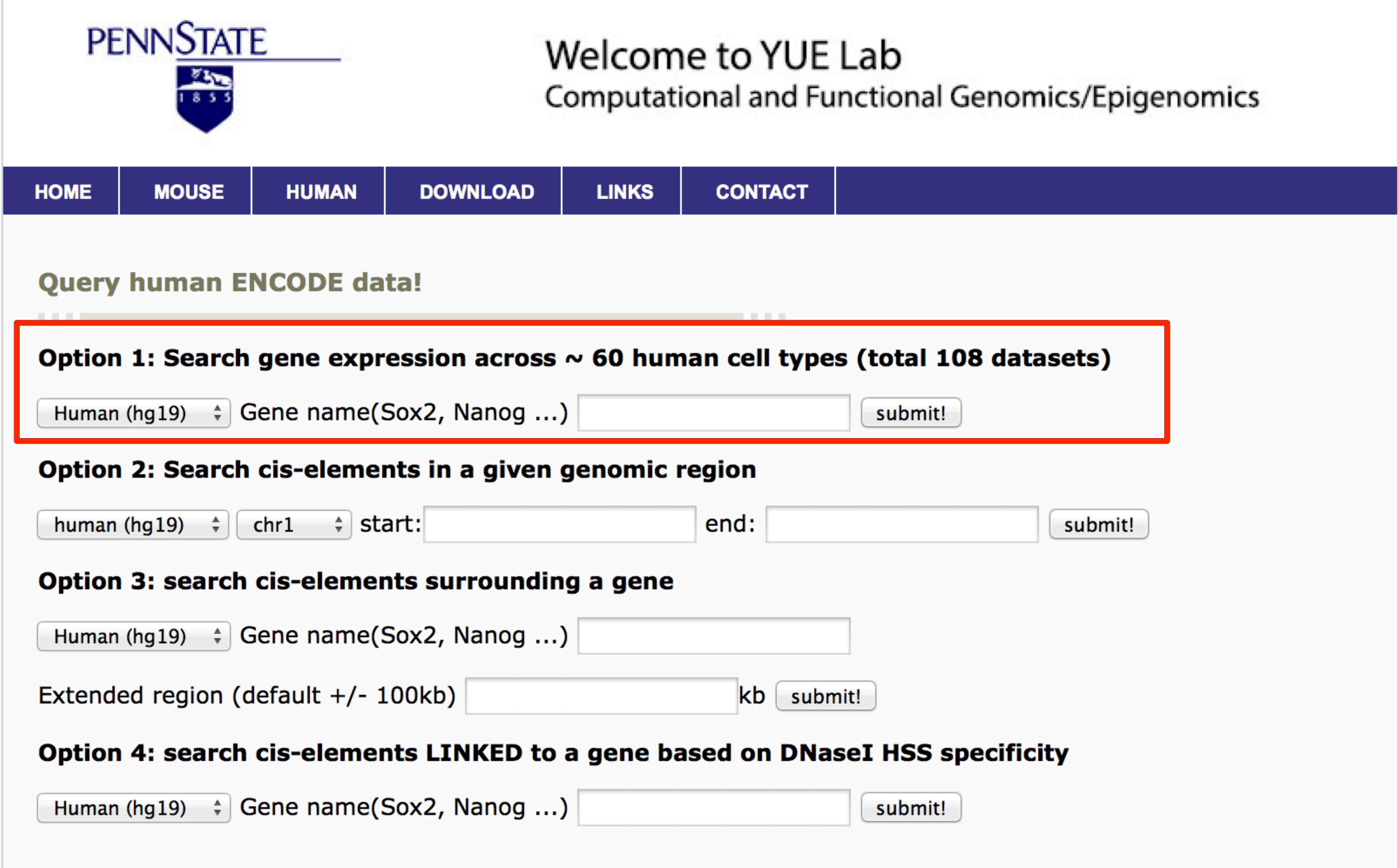
Open chromatin (DNase-seq)

DNase I hypersensitive sites (also known as DNase-seq peaks) computed from ~300 human and mouse experiments.

[Raw Data | Peaks]

CTCF DHS Profile

Click human tab, and search for the expression for Sox2



PENNSTATE

1855

Welcome to YUE Lab

Computational and Functional Genomics/Epigenomics

Query human ENCODE data!

Option 1: Search gene expression across ~ 60 human cell types (total 108 datasets)

Human (hg19) Gene name(Sox2, Nanog ...) submit!

Option 2: Search cis-elements in a given genomic region

human (hg19) chr1 start: end: submit!

Option 3: search cis-elements surrounding a gene

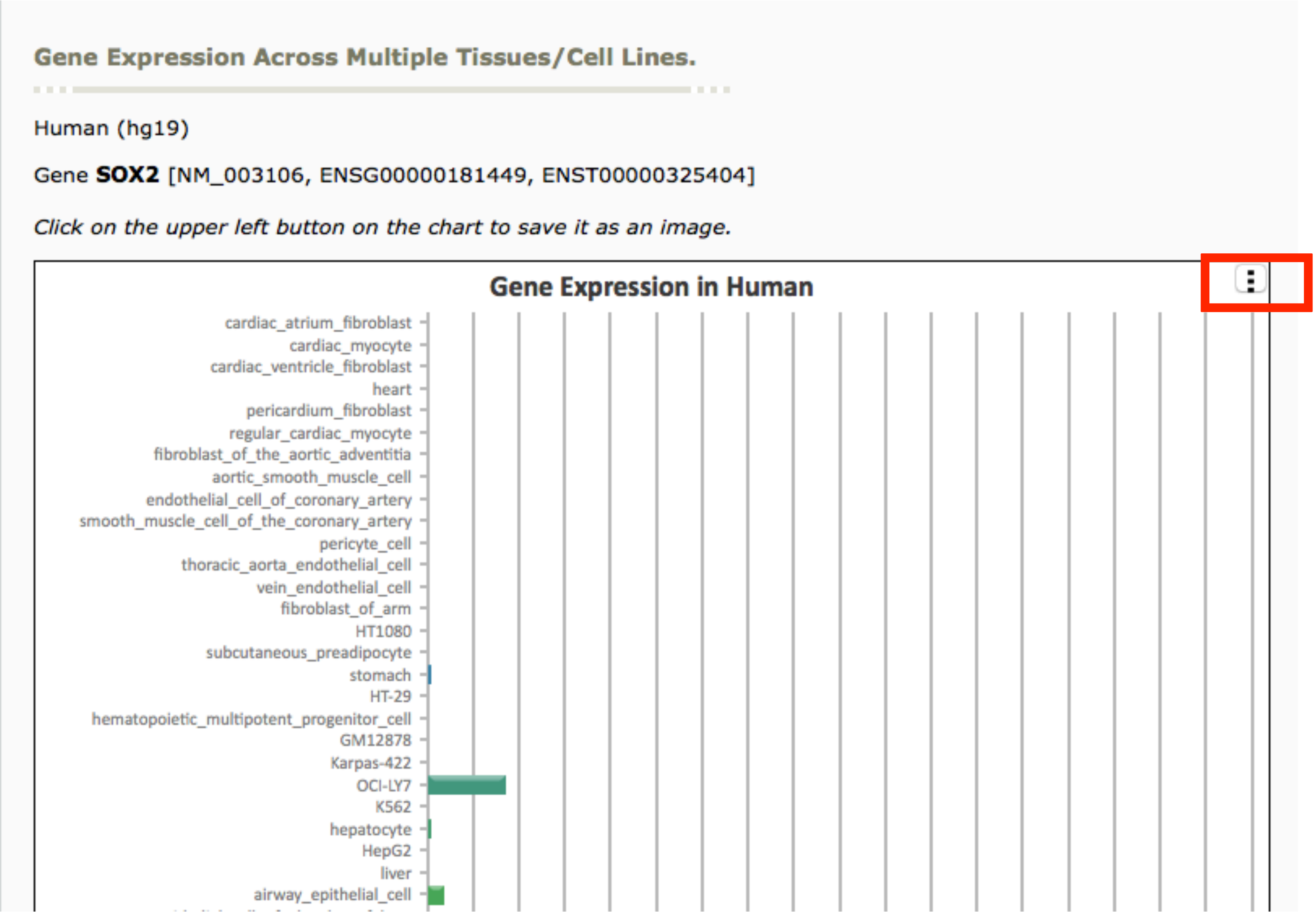
Human (hg19) Gene name(Sox2, Nanog ...)

Extended region (default +/- 100kb) kb submit!

Option 4: search cis-elements LINKED to a gene based on DNaseI HSS specificity

Human (hg19) Gene name(Sox2, Nanog ...) submit!

Visualize and save the gene expression profile



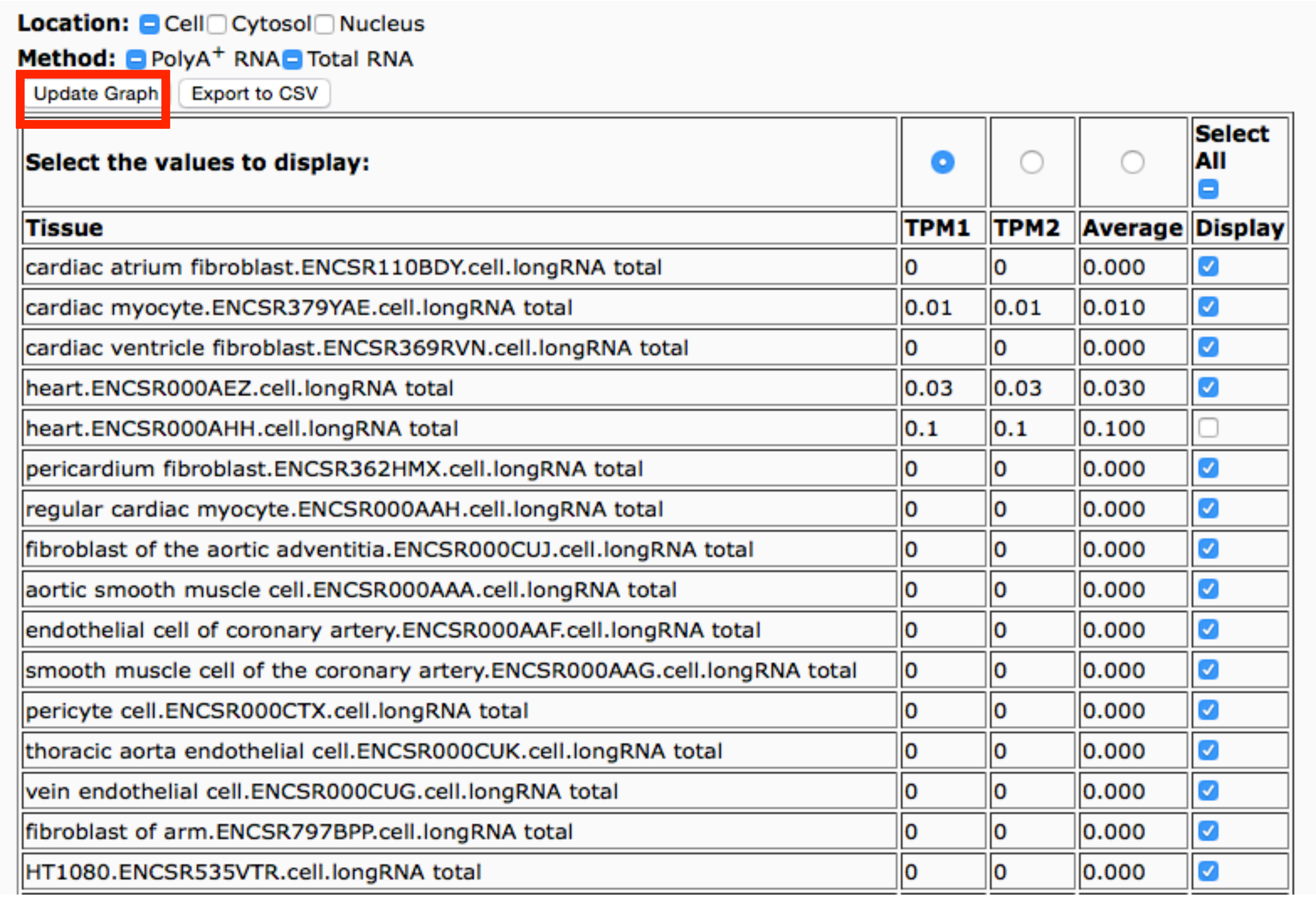
Gene Expression Across Multiple Tissues/Cell Lines.

Human (hg19)

Gene **SOX2** [NM_003106, ENSG00000181449, ENST00000325404]

Click on the upper left button on the chart to save it as an image.

You can choose the cell types to display!

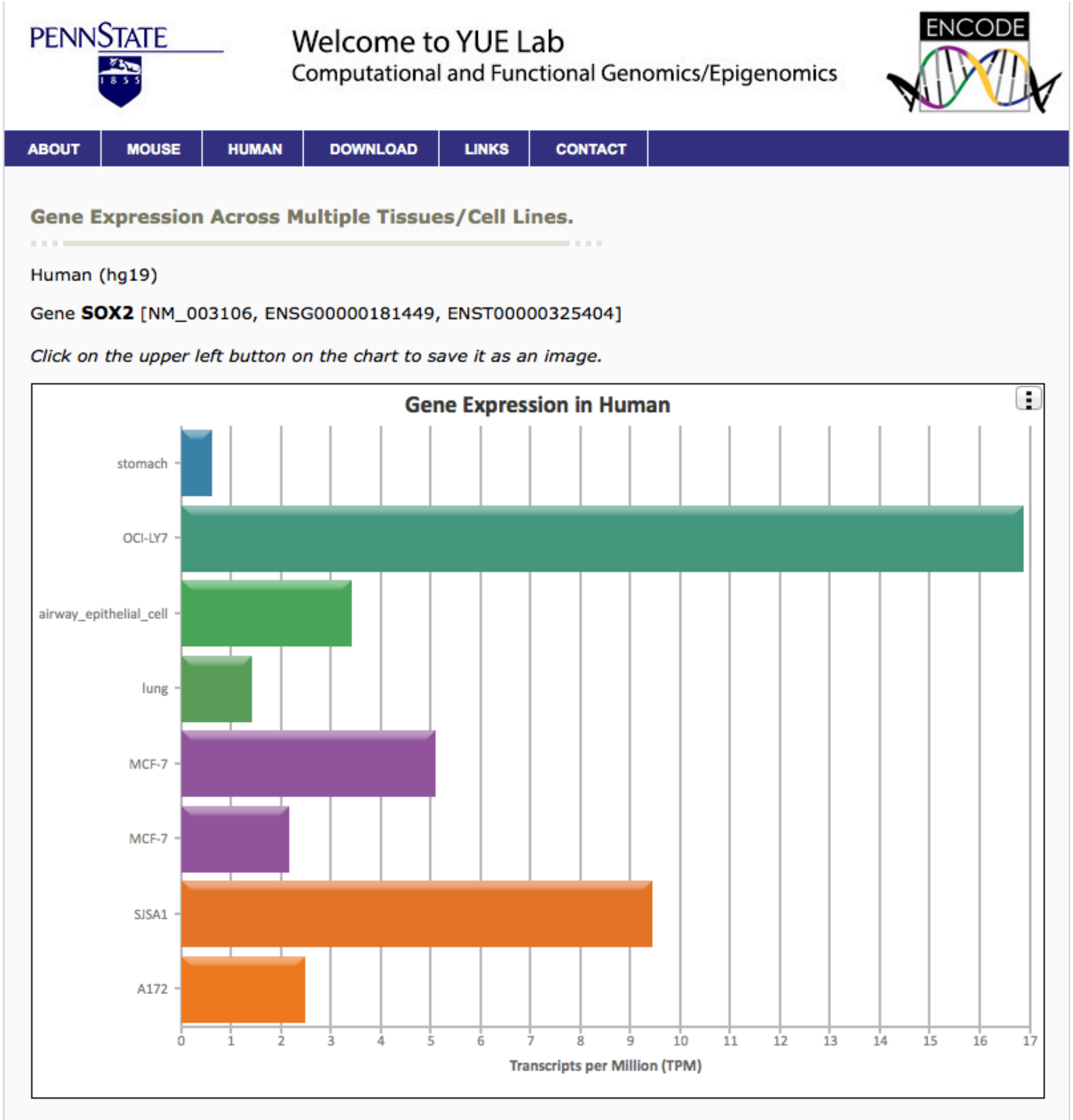


Location: ⊟ Cell ☐ Cytosol ☐ Nucleus

Method: ⊟ PolyA$^+$ RNA ⊟ Total RNA

Select the values to display:	●	○	○	**Select All** ⊟
Tissue	**TPM1**	**TPM2**	**Average**	**Display**
cardiac atrium fibroblast.ENCSR110BDY.cell.longRNA total	0	0	0.000	☑
cardiac myocyte.ENCSR379YAE.cell.longRNA total	0.01	0.01	0.010	☑
cardiac ventricle fibroblast.ENCSR369RVN.cell.longRNA total	0	0	0.000	☑
heart.ENCSR000AEZ.cell.longRNA total	0.03	0.03	0.030	☑
heart.ENCSR000AHH.cell.longRNA total	0.1	0.1	0.100	☐
pericardium fibroblast.ENCSR362HMX.cell.longRNA total	0	0	0.000	☑
regular cardiac myocyte.ENCSR000AAH.cell.longRNA total	0	0	0.000	☑
fibroblast of the aortic adventitia.ENCSR000CUJ.cell.longRNA total	0	0	0.000	☑
aortic smooth muscle cell.ENCSR000AAA.cell.longRNA total	0	0	0.000	☑
endothelial cell of coronary artery.ENCSR000AAF.cell.longRNA total	0	0	0.000	☑
smooth muscle cell of the coronary artery.ENCSR000AAG.cell.longRNA total	0	0	0.000	☑
pericyte cell.ENCSR000CTX.cell.longRNA total	0	0	0.000	☑
thoracic aorta endothelial cell.ENCSR000CUK.cell.longRNA total	0	0	0.000	☑
vein endothelial cell.ENCSR000CUG.cell.longRNA total	0	0	0.000	☑
fibroblast of arm.ENCSR797BPP.cell.longRNA total	0	0	0.000	☑
HT1080.ENCSR535VTR.cell.longRNA total	0	0	0.000	☑

You can choose the cell types to display!



Welcome to YUE Lab

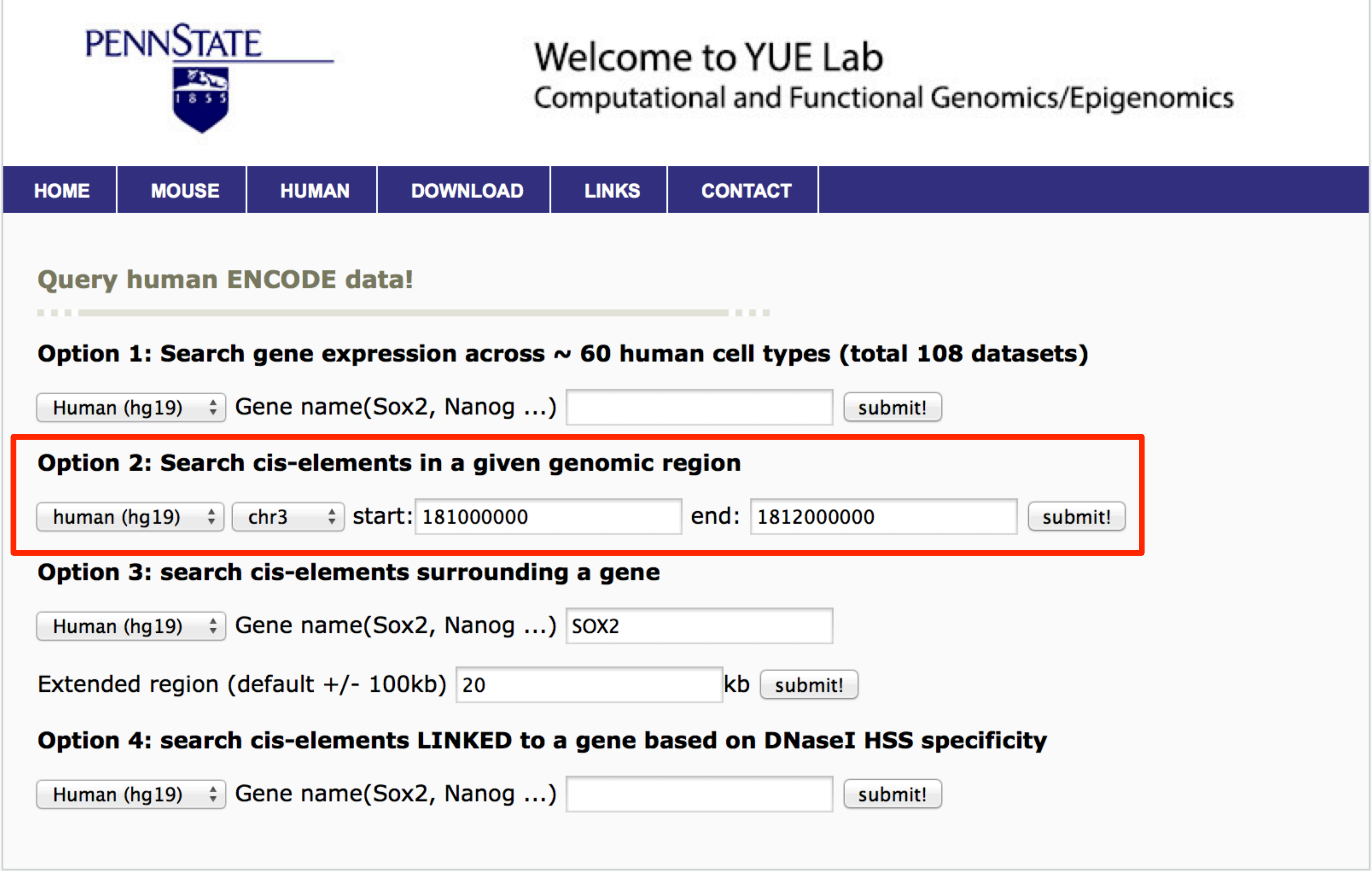
Computational and Functional Genomics/Epigenomics

Gene Expression Across Multiple Tissues/Cell Lines.

Human (hg19)

Gene **SOX2** [NM_003106, ENSG00000181449, ENST00000325404]

Click on the upper left button on the chart to save it as an image.



Welcome to YUE Lab

Computational and Functional Genomics/Epigenomics

Query human ENCODE data!

Option 1: Search gene expression across ~ 60 human cell types (total 108 datasets)

Human (hg19) Gene name(Sox2, Nanog ...) submit!

Option 2: Search cis-elements in a given genomic region

human (hg19) chr3 start: 181000000 end: 1812000000 submit!

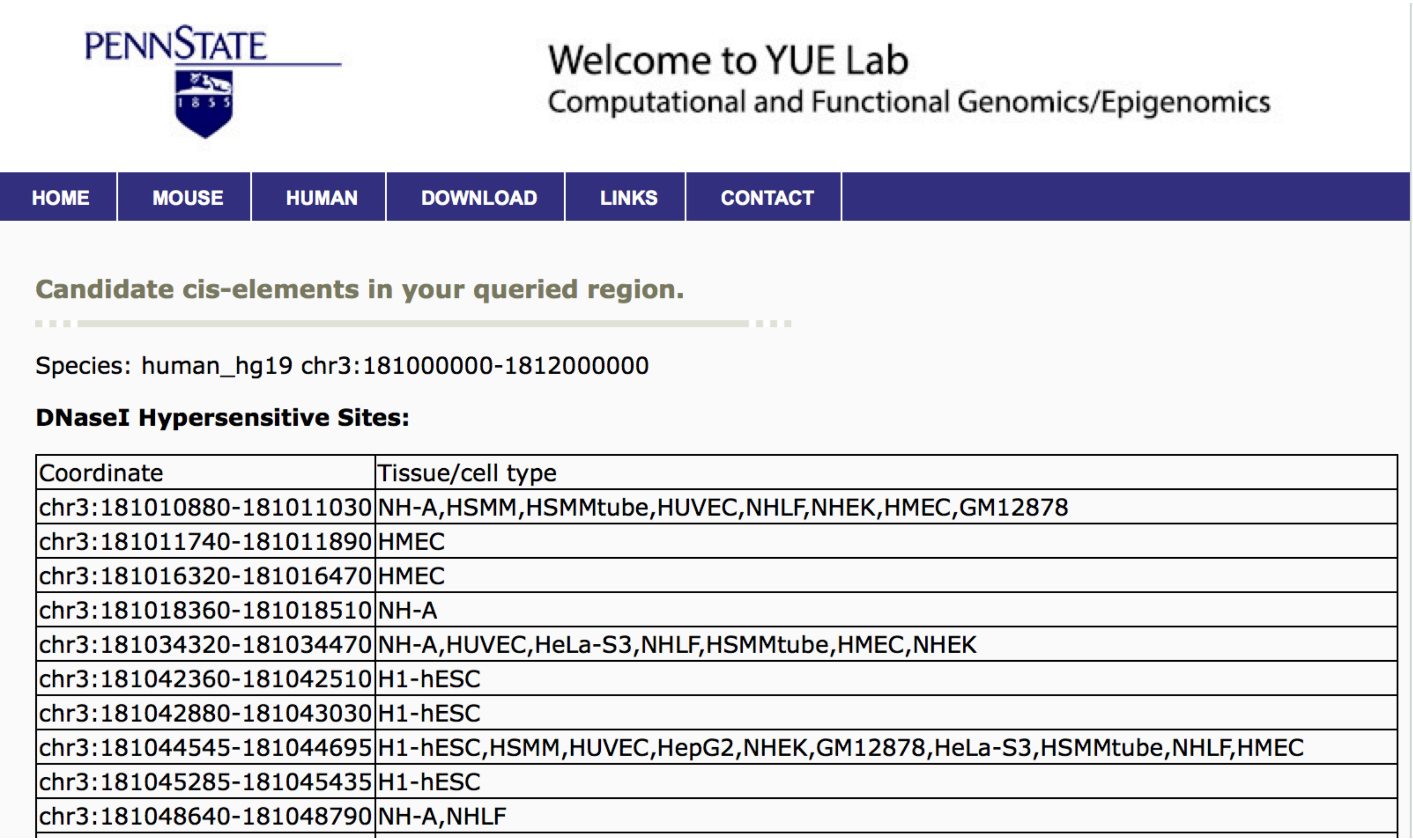
Option 3: search cis-elements surrounding a gene

Human (hg19) Gene name(Sox2, Nanog ...) SOX2

Extended region (default +/- 100kb) 20 kb submit!

Option 4: search cis-elements LINKED to a gene based on DNaseI HSS specificity

Human (hg19) Gene name(Sox2, Nanog ...) submit!



Welcome to YUE Lab

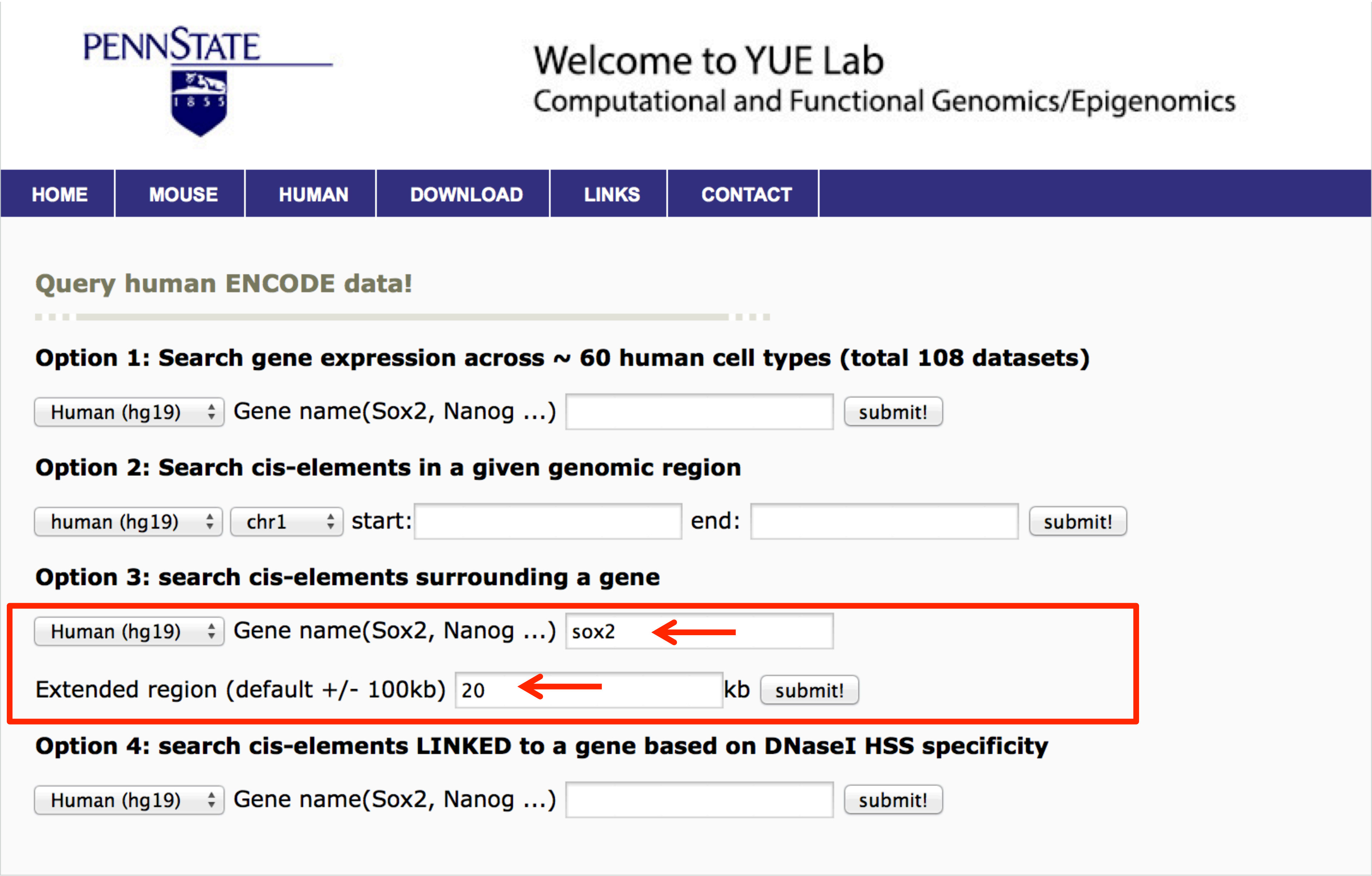
Computational and Functional Genomics/Epigenomics

Candidate cis-elements in your queried region.

Species: human_hg19 chr3:181000000-1812000000

DNaseI Hypersensitive Sites:

Coordinate	Tissue/cell type
chr3:181010880-181011030	NH-A,HSMM,HSMMtube,HUVEC,NHLF,NHEK,HMEC,GM12878
chr3:181011740-181011890	HMEC
chr3:181016320-181016470	HMEC
chr3:181018360-181018510	NH-A
chr3:181034320-181034470	NH-A,HUVEC,HeLa-S3,NHLF,HSMMtube,HMEC,NHEK
chr3:181042360-181042510	H1-hESC
chr3:181042880-181043030	H1-hESC
chr3:181044545-181044695	H1-hESC,HSMM,HUVEC,HepG2,NHEK,GM12878,HeLa-S3,HSMMtube,NHLF,HMEC
chr3:181045285-181045435	H1-hESC
chr3:181048640-181048790	NH-A,NHLF



Welcome to YUE Lab
Computational and Functional Genomics/Epigenomics

Query human ENCODE data!

Option 1: Search gene expression across ~ 60 human cell types (total 108 datasets)

Human (hg19) Gene name(Sox2, Nanog ...) submit!

Option 2: Search cis-elements in a given genomic region

human (hg19) chr1 start: end: submit!

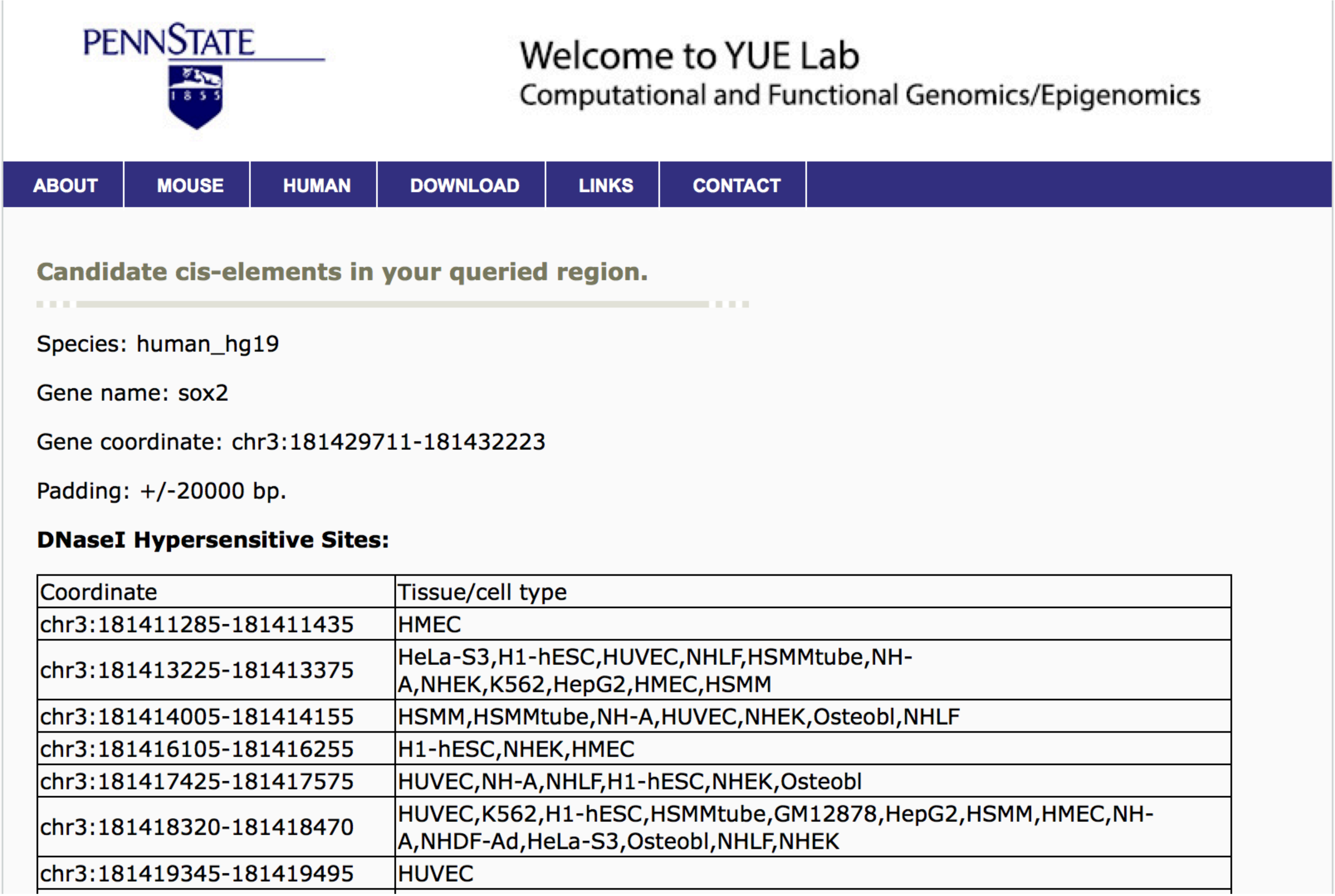
Option 3: search cis-elements surrounding a gene

Human (hg19) Gene name(Sox2, Nanog ...) sox2

Extended region (default +/- 100kb) 20 kb submit!

Option 4: search cis-elements LINKED to a gene based on DNaseI HSS specificity

Human (hg19) Gene name(Sox2, Nanog ...) submit!



Welcome to YUE Lab

Computational and Functional Genomics/Epigenomics

Candidate cis-elements in your queried region.

Species: human_hg19

Gene name: sox2

Gene coordinate: chr3:181429711-181432223

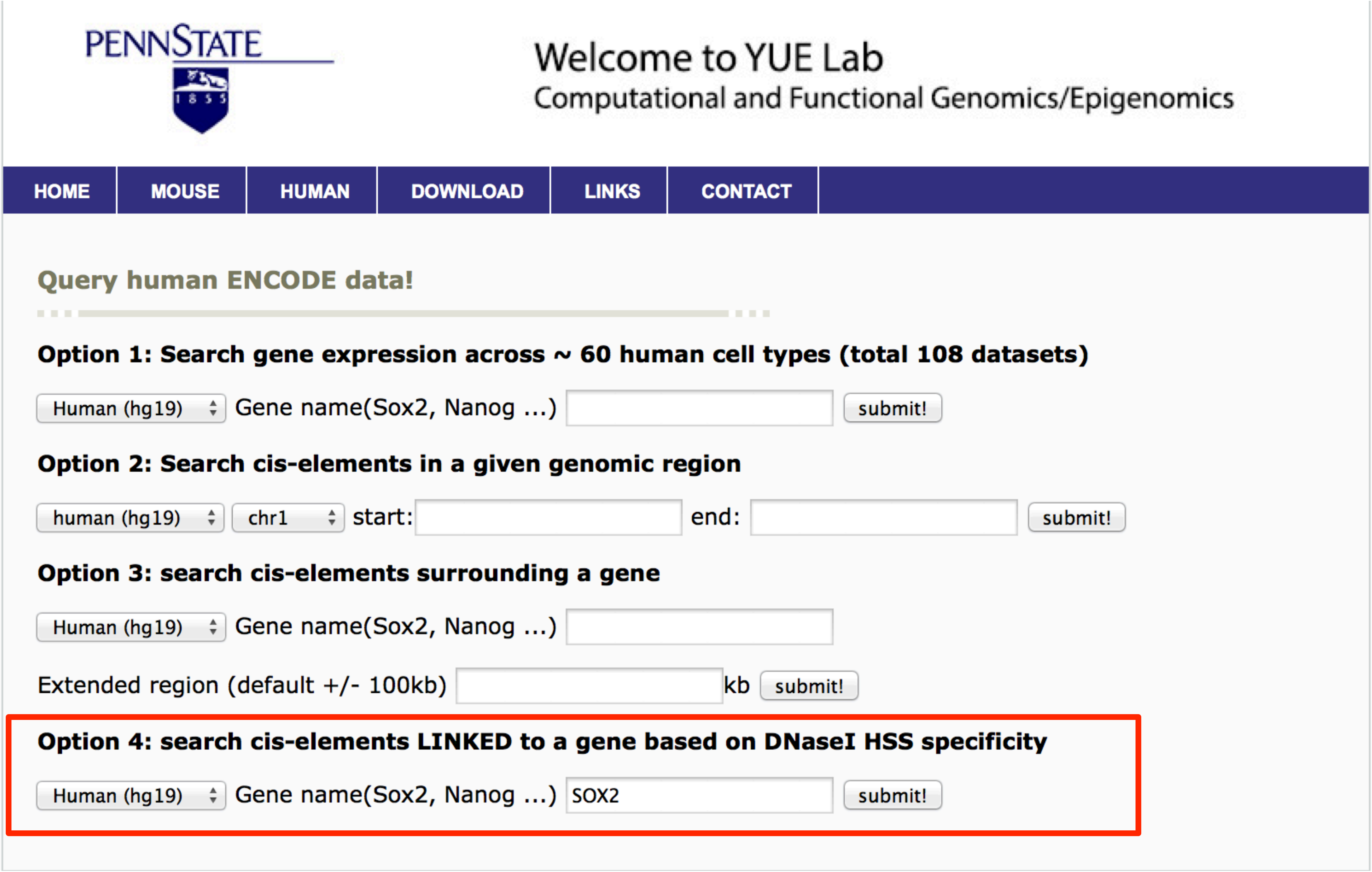
Padding: +/-20000 bp.

DNaseI Hypersensitive Sites:

Coordinate	Tissue/cell type
chr3:181411285-181411435	HMEC
chr3:181413225-181413375	HeLa-S3,H1-hESC,HUVEC,NHLF,HSMMtube,NH-A,NHEK,K562,HepG2,HMEC,HSMM
chr3:181414005-181414155	HSMM,HSMMtube,NH-A,HUVEC,NHEK,Osteobl,NHLF
chr3:181416105-181416255	H1-hESC,NHEK,HMEC
chr3:181417425-181417575	HUVEC,NH-A,NHLF,H1-hESC,NHEK,Osteobl
chr3:181418320-181418470	HUVEC,K562,H1-hESC,HSMMtube,GM12878,HepG2,HSMM,HMEC,NH-A,NHDF-Ad,HeLa-S3,Osteobl,NHLF,NHEK
chr3:181419345-181419495	HUVEC

TF binding Site:

Coordinate	TF	Bindings in tissues
chr3:181413225-181413375	multiple	BACH1(H1-hESC), CHD1(H1-hESC), CTBP2(H1-hESC), CTCF(AG09319), CTCF(AoAF), CTCF(Fibrobl), CTCF(GM12874), CTCF(GM12892), CTCF(GM19238), CTCF(GM19239), CTCF(GM19240), CTCF(Gliobla), CTCF(H1-hESC), CTCF(HA-sp), CTCF(HAc), CTCF(HBMEC), CTCF(HCM), CTCF(HMEC), CTCF(HRPEpiC), CTCF(HUVEC), CTCF(K562), CTCF(NHEK), CTCF(NHLF), CTCF(SAEC), E2F6(K562), EZH2(NHDF-Ad), HDAC2(H1-hESC), MAX(K562), POLR2A(HUVEC), RAD21(H1-hESC), RBBP5(H1-hESC), ZBTB7A(K562)
chr3:181414005-181414155	EZH2	EZH2(H1-hESC)
chr3:181416105-181416255	BATF	BATF(GM12878)
chr3:181417425-181417575	JUND, SP1, TBP	JUND(H1-hESC), SP1(H1-hESC), TBP(H1-hESC)
chr3:181418320-181418470	multiple	CEBPB(A549), CEBPB(HeLa-S3), CEBPB(K562), CTCF(A549), CTCF(AG04449), CTCF(AG04450), CTCF(AG09309), CTCF(AG09319), CTCF(AG10803), CTCF(AoAF), CTCF(BE2_C), CTCF(BJ), CTCF(Caco-2), CTCF(Dnd41), CTCF(ECC-1), CTCF(Fibrobl), CTCF(GM06990), CTCF(GM12864), CTCF(GM12865), CTCF(GM12872), CTCF(GM12873), CTCF(GM12874), CTCF(GM12875), CTCF(GM12878), CTCF(GM12891), CTCF(GM12892), CTCF(GM19238), CTCF(GM19239), CTCF(GM19240), CTCF(Gliobla), CTCF(H1-hESC), CTCF(HA-sp), CTCF(HAc), CTCF(HBMEC), CTCF(HCFaa), CTCF(HCM), CTCF(HCPEpiC), CTCF(HCT-116), CTCF(HEEpiC), CTCF(HEK293), CTCF(HFF), CTCF(HFF-Myc), CTCF(HL-60), CTCF(HMEC), CTCF(HMF), CTCF(HPAF), CTCF(HPF), CTCF(HRE), CTCF(HRPEpiC), CTCF(HSMM), CTCF(HSMMtube),



Welcome to YUE Lab

Computational and Functional Genomics/Epigenomics

Query human ENCODE data!

Option 1: Search gene expression across ~ 60 human cell types (total 108 datasets)

Human (hg19) Gene name(Sox2, Nanog ...) submit!

Option 2: Search cis-elements in a given genomic region

human (hg19) chr1 start: end: submit!

Option 3: search cis-elements surrounding a gene

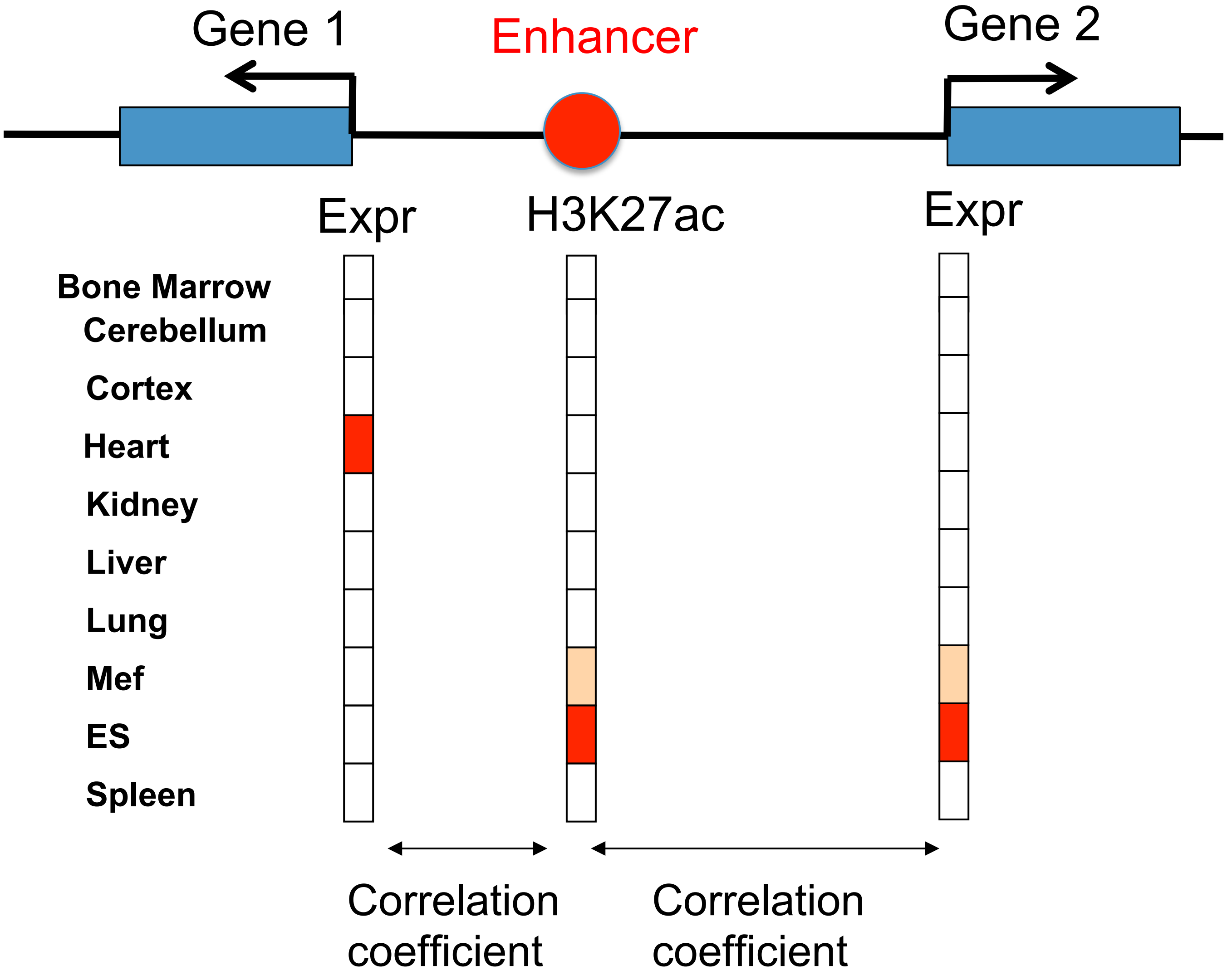
Human (hg19) Gene name(Sox2, Nanog ...)

Extended region (default +/- 100kb) kb submit!

Option 4: search cis-elements LINKED to a gene based on DNaseI HSS specificity

Human (hg19) Gene name(Sox2, Nanog ...) SOX2 submit!

Linking enhancers to target promoters



Gene 1

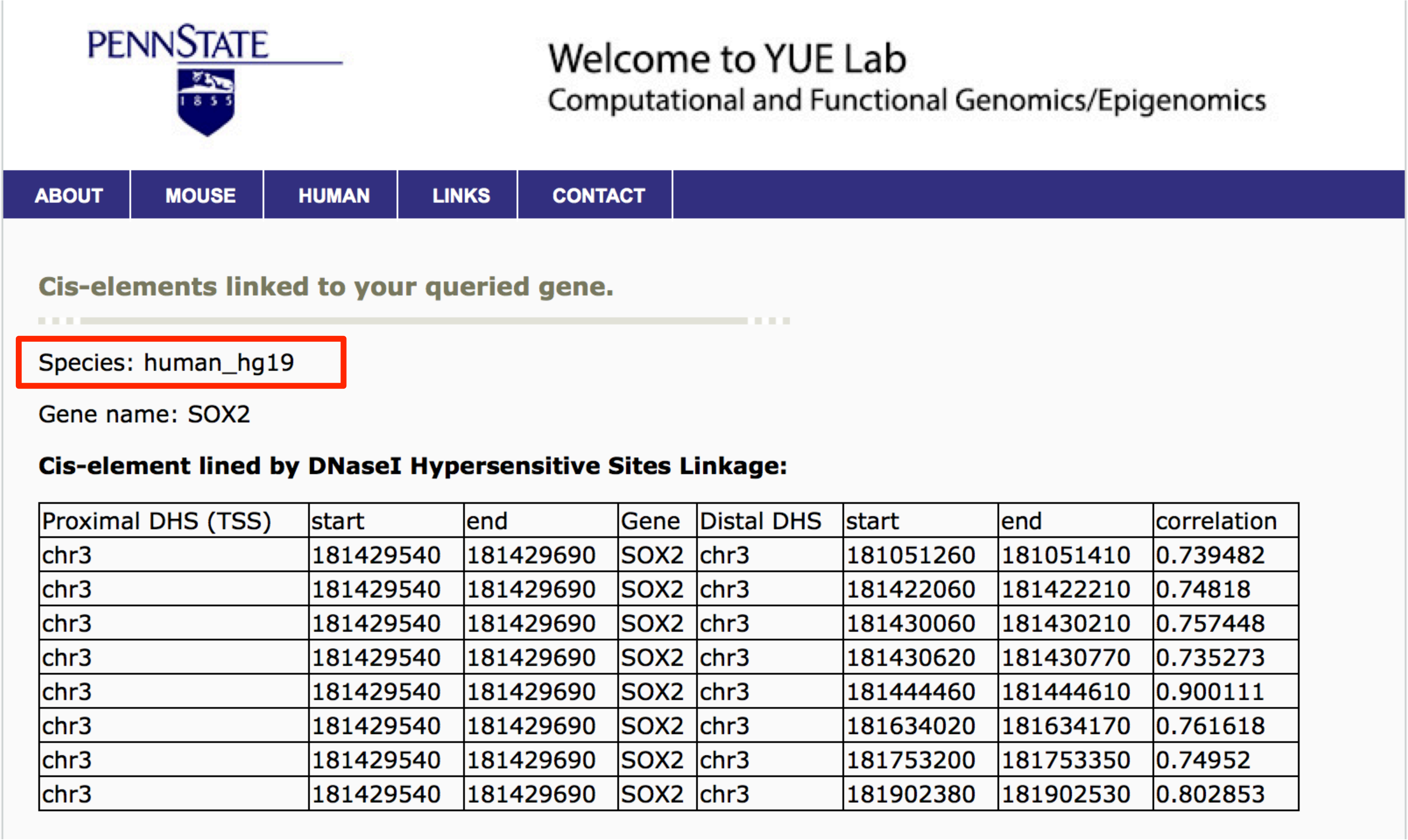
Enhancer

Gene 2

	Expr	H3K27ac	Expr
Bone Marrow			
Cerebellum			
Cortex			
Heart	🟥		
Kidney			
Liver			
Lung			
Mef		🟧	🟧
ES		🟥	🟥
Spleen			

Correlation coefficient

Correlation coefficient



Welcome to YUE Lab

Computational and Functional Genomics/Epigenomics

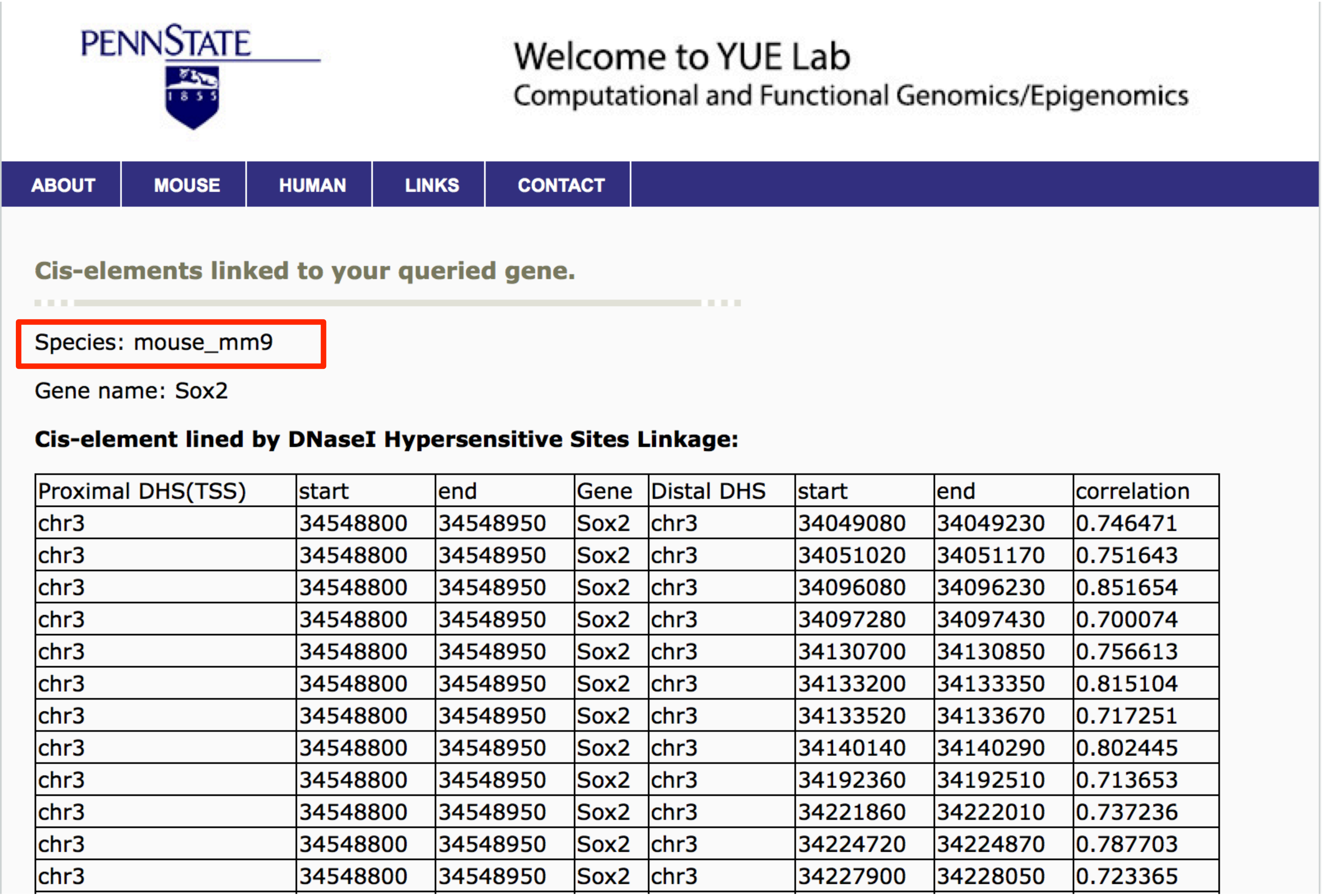
Cis-elements linked to your queried gene.

Species: human_hg19

Gene name: SOX2

Cis-element lined by DNaseI Hypersensitive Sites Linkage:

Proximal DHS (TSS)	start	end	Gene	Distal DHS	start	end	correlation
chr3	181429540	181429690	SOX2	chr3	181051260	181051410	0.739482
chr3	181429540	181429690	SOX2	chr3	181422060	181422210	0.74818
chr3	181429540	181429690	SOX2	chr3	181430060	181430210	0.757448
chr3	181429540	181429690	SOX2	chr3	181430620	181430770	0.735273
chr3	181429540	181429690	SOX2	chr3	181444460	181444610	0.900111
chr3	181429540	181429690	SOX2	chr3	181634020	181634170	0.761618
chr3	181429540	181429690	SOX2	chr3	181753200	181753350	0.74952
chr3	181429540	181429690	SOX2	chr3	181902380	181902530	0.802853



Welcome to YUE Lab

Computational and Functional Genomics/Epigenomics

Cis-elements linked to your queried gene.

Species: mouse_mm9

Gene name: Sox2

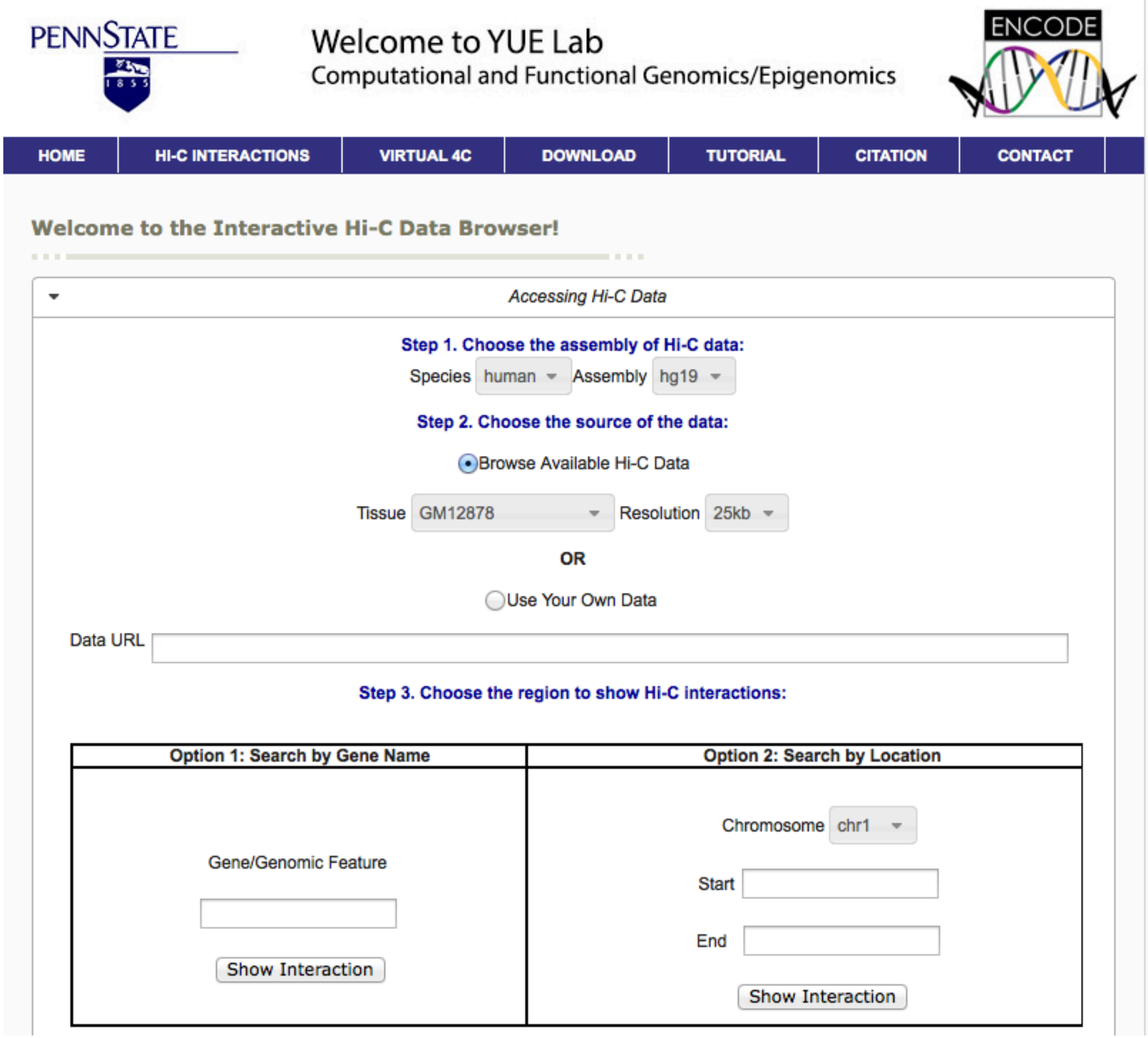
Cis-element lined by DNaseI Hypersensitive Sites Linkage:

Proximal DHS(TSS)	start	end	Gene	Distal DHS	start	end	correlation
chr3	34548800	34548950	Sox2	chr3	34049080	34049230	0.746471
chr3	34548800	34548950	Sox2	chr3	34051020	34051170	0.751643
chr3	34548800	34548950	Sox2	chr3	34096080	34096230	0.851654
chr3	34548800	34548950	Sox2	chr3	34097280	34097430	0.700074
chr3	34548800	34548950	Sox2	chr3	34130700	34130850	0.756613
chr3	34548800	34548950	Sox2	chr3	34133200	34133350	0.815104
chr3	34548800	34548950	Sox2	chr3	34133520	34133670	0.717251
chr3	34548800	34548950	Sox2	chr3	34140140	34140290	0.802445
chr3	34548800	34548950	Sox2	chr3	34192360	34192510	0.713653
chr3	34548800	34548950	Sox2	chr3	34221860	34222010	0.737236
chr3	34548800	34548950	Sox2	chr3	34224720	34224870	0.787703
chr3	34548800	34548950	Sox2	chr3	34227900	34228050	0.723365

Outline for the workshop

1. **Brief tour of ENCODE project;**
2. **Search and download through ENCODE portal**
3. **Search and visualize ENCODE annotations (gene expression, promoters, enhancers …)**
4. **Visualize ENCODE chromatin interaction data (Hi-C, ChIA-PET)**

http://3dgenome.org



Welcome to YUE Lab

Computational and Functional Genomics/Epigenomics

Welcome to the Interactive Hi-C Data Browser!

Accessing Hi-C Data

Step 1. Choose the assembly of Hi-C data:

Species: human | Assembly: hg19

Step 2. Choose the source of the data:

● Browse Available Hi-C Data

Tissue: GM12878 | Resolution: 25kb

OR

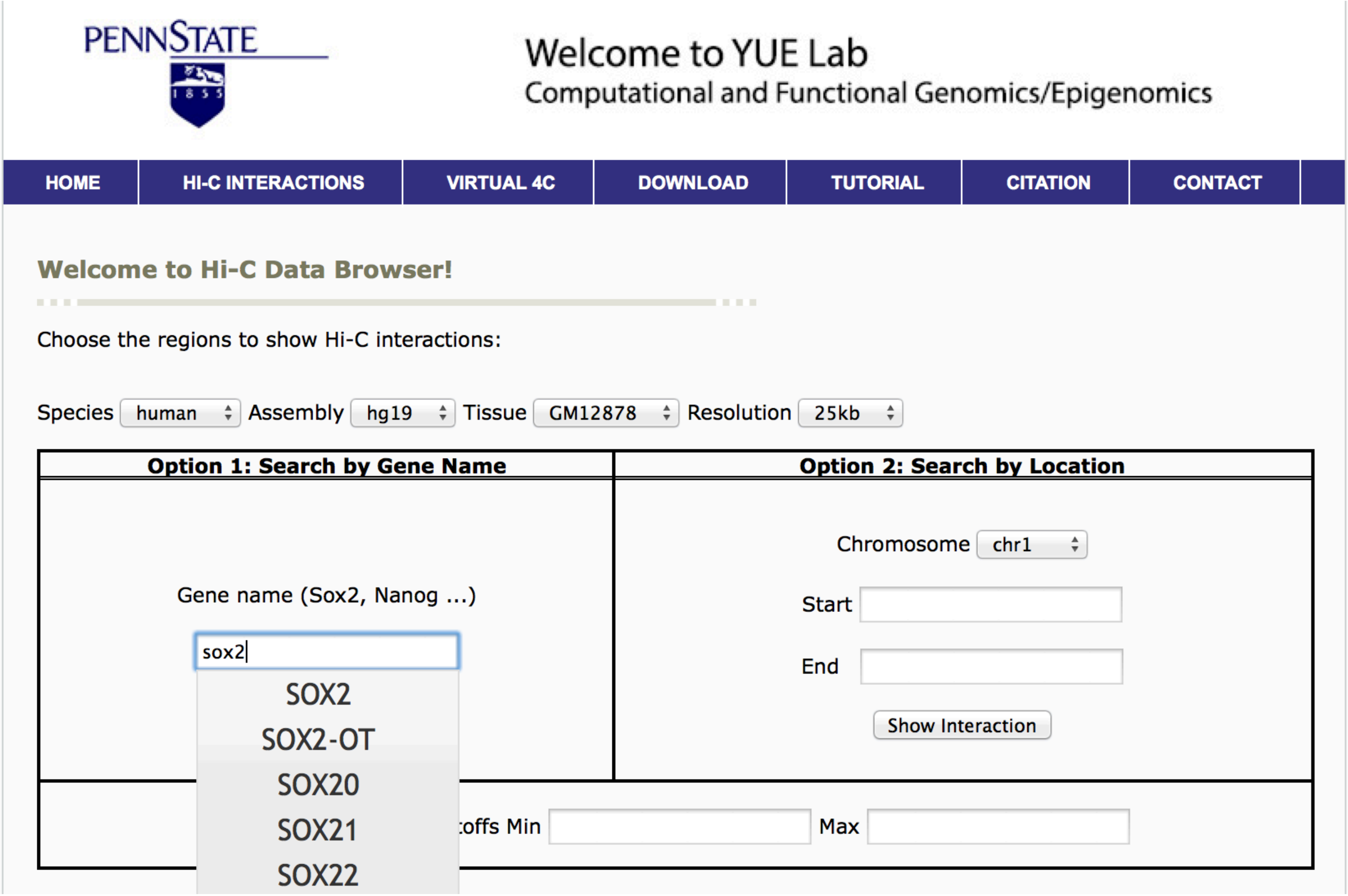
○ Use Your Own Data

Data URL:

Step 3. Choose the region to show Hi-C interactions:

Option 1: Search by Gene Name	Option 2: Search by Location
Gene/Genomic Feature: Show Interaction	Chromosome: chr1 Start: End: Show Interaction

Users can query by gene name or genomic loci



Welcome to Hi-C Data Browser!

Choose the regions to show Hi-C interactions:

Species: human
Assembly: hg19
Tissue: GM12878
Resolution: 25kb

Option 1: Search by Gene Name

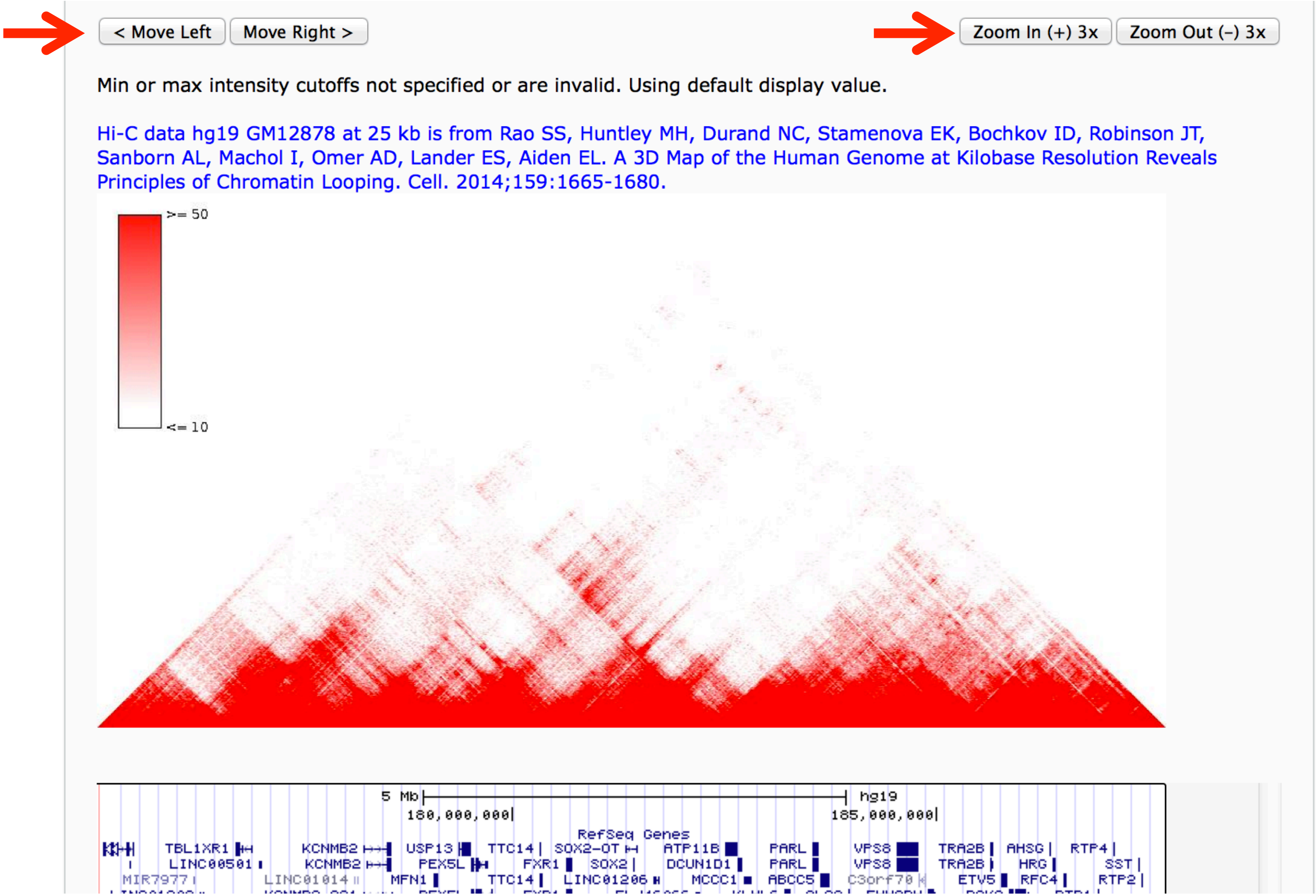
Gene name (Sox2, Nanog ...): sox2

Option 2: Search by Location

Chromosome: chr1
Start:
End:

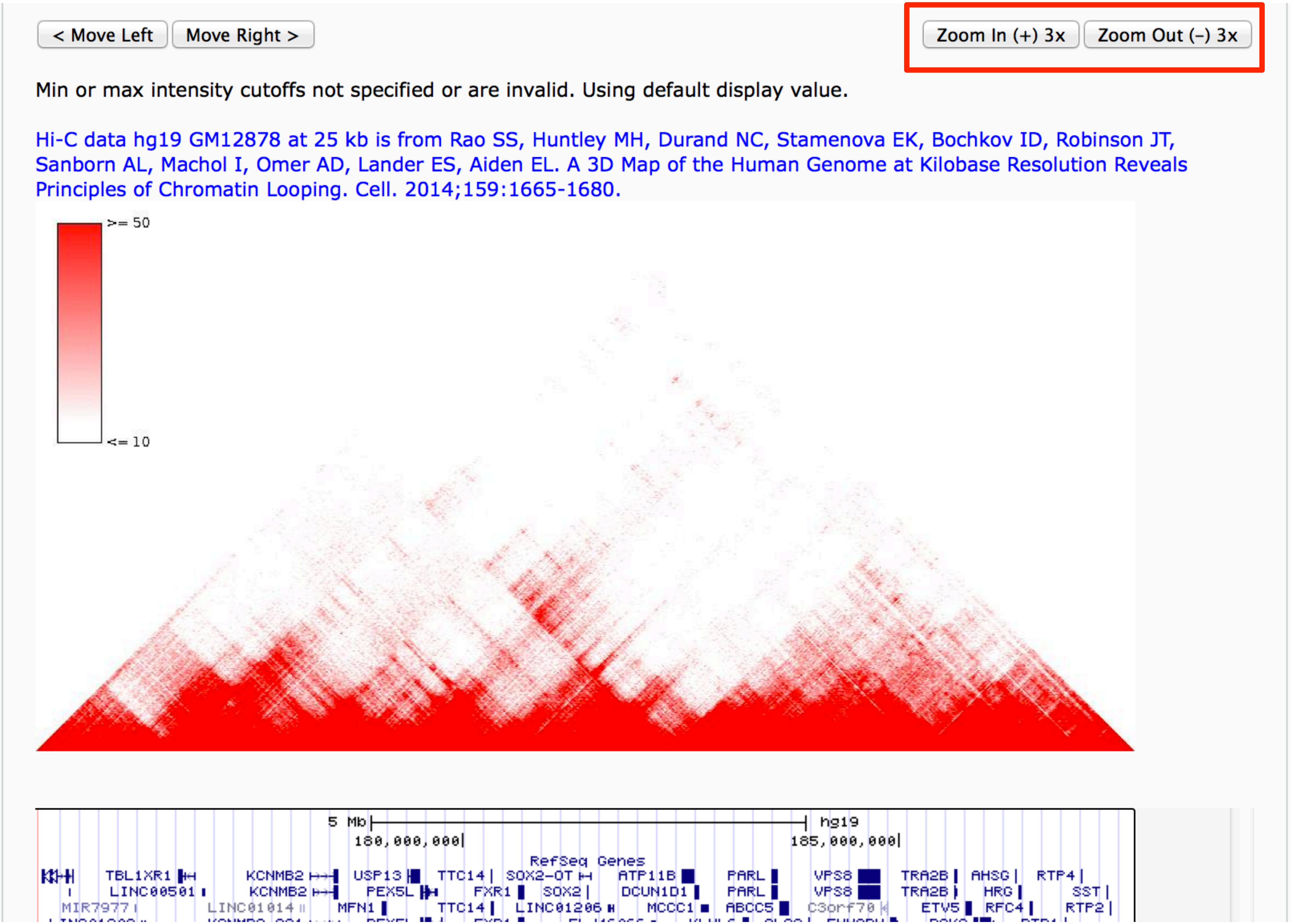
offs Min
Max

Navigation bars to explore Hi-C data



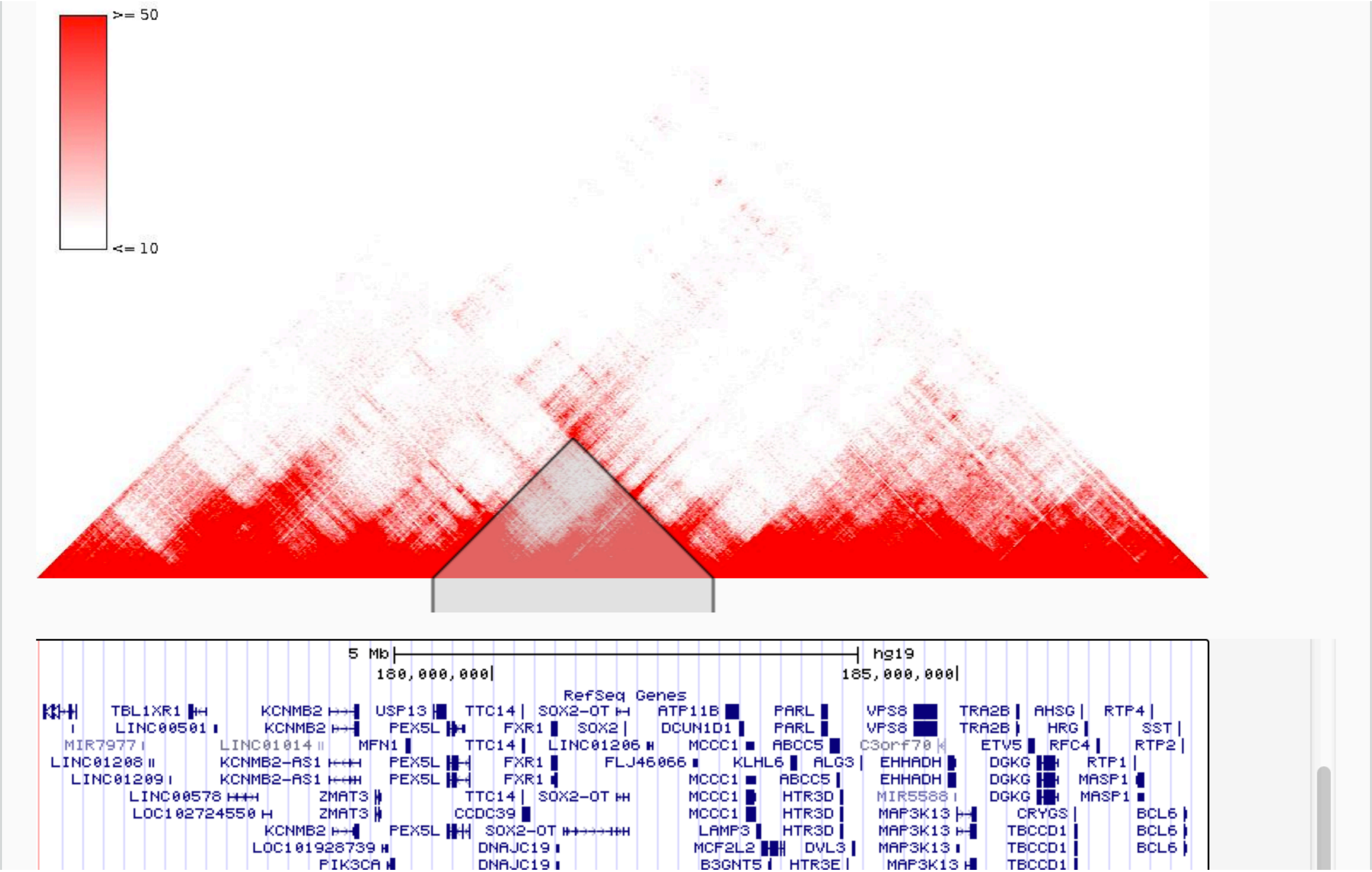
Min or max intensity cutoffs not specified or are invalid. Using default display value.

Hi-C data hg19 GM12878 at 25 kb is from Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014;159:1665-1680.

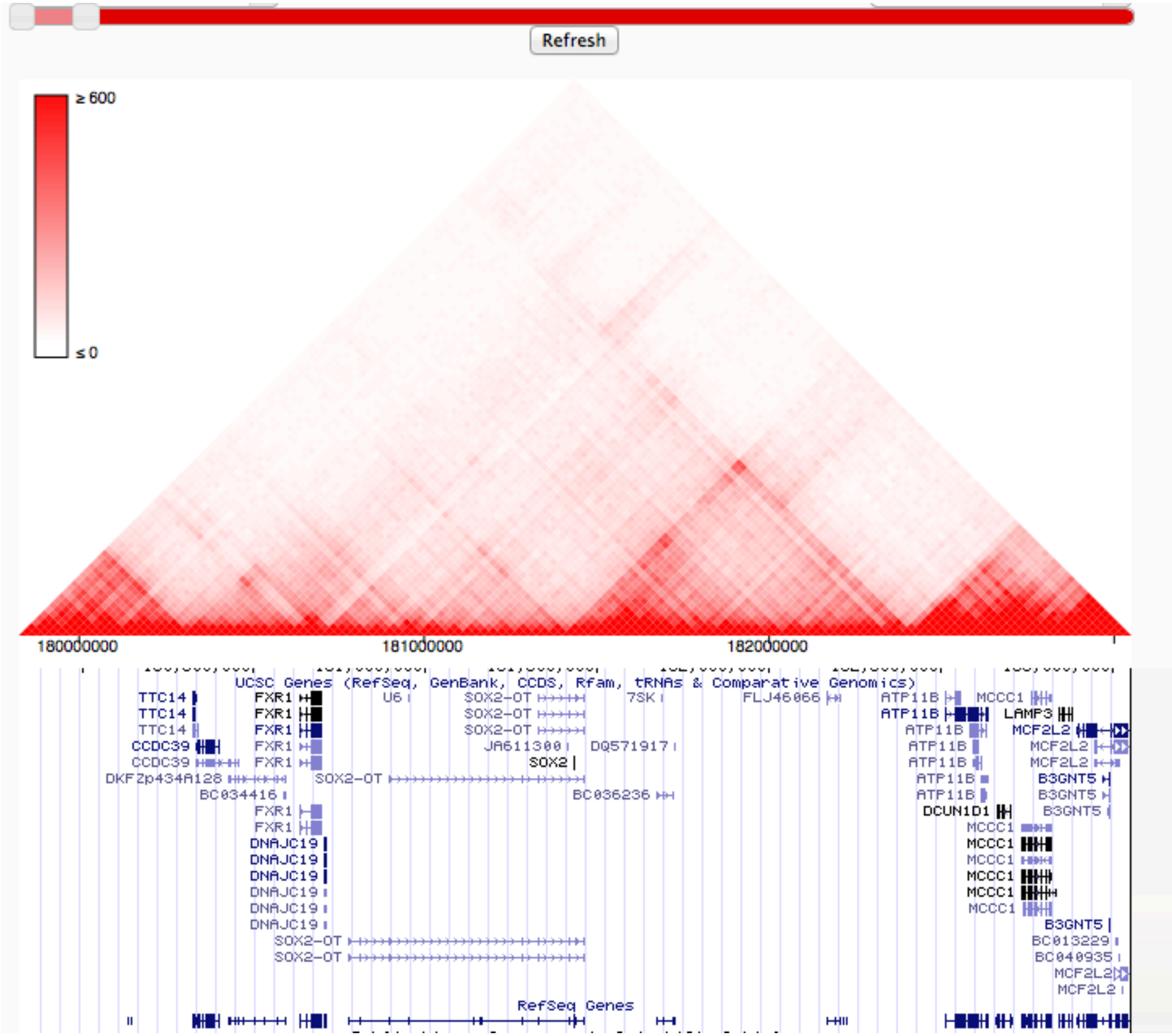


Min or max intensity cutoffs not specified or are invalid. Using default display value.

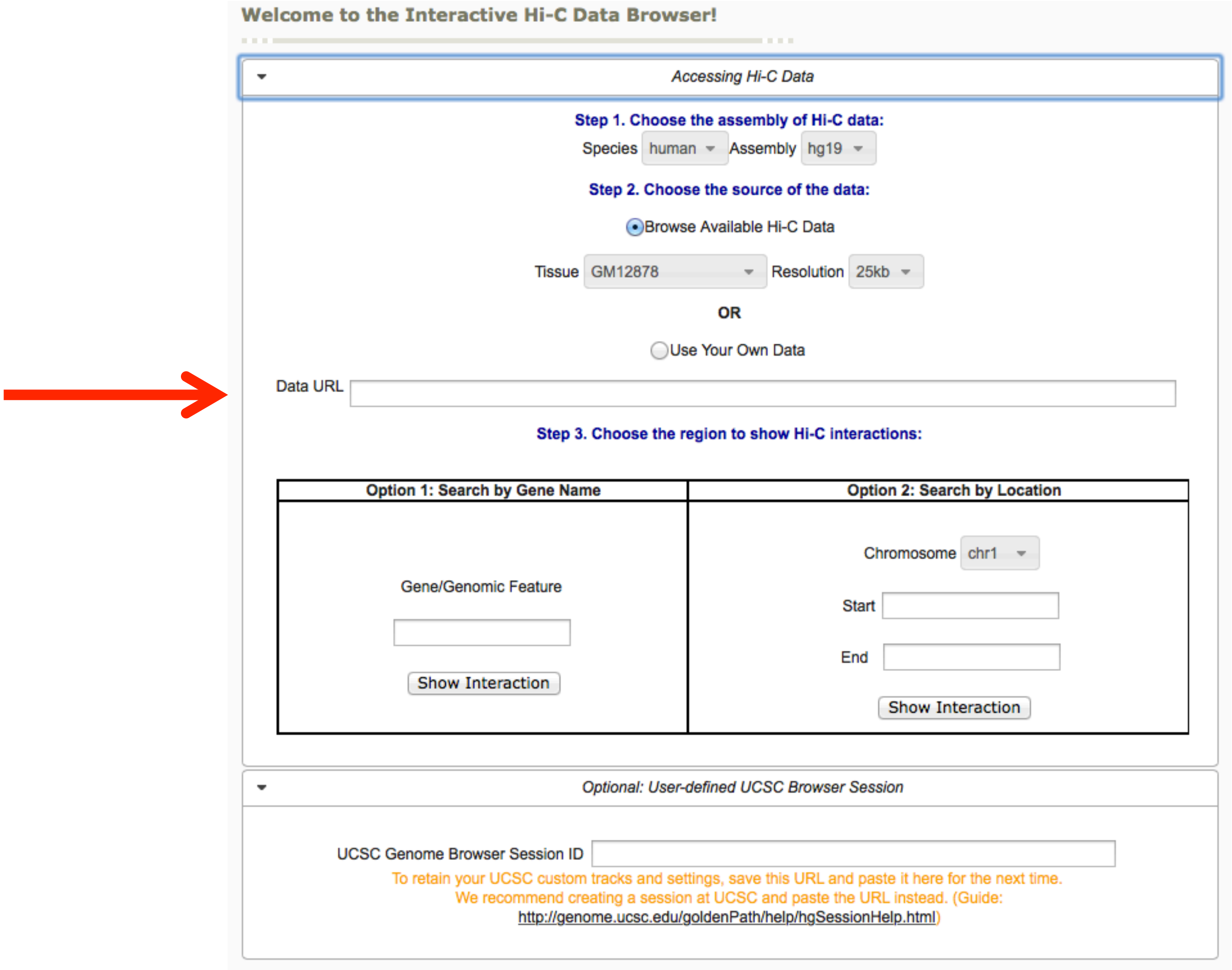
Hi-C data hg19 GM12878 at 25 kb is from Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014;159:1665-1680.


>= 50
<= 10
5 Mb
hg19
180,000,000
185,000,000
RefSeq Genes
TBL1XR1
LINC00501
MIR7977
LINC01208
LINC01209
LINC00578
LOC102724550
KCNMB2
KCNMB2
LINC01014
KCNMB2-AS1
KCNMB2-AS1
ZMAT3
ZMAT3
KCNMB2
LOC101928739
PIK3CA
USP13
PEX5L
MFN1
PEX5L
PEX5L
PEX5L
TTC14
FXR1
TTC14
FXR1
FXR1
TTC14
CCDC39
SOX2-OT
DNAJC19
DNAJC19
SOX2-OT
SOX2
LINC01206
FLJ46066
SOX2-OT
ATP11B
DCUN1D1
MCCC1
KLHL6
MCCC1
MCCC1
MCCC1
LAMP3
MCF2L2
B3GNT5
PARL
PARL
ABCC5
ALG3
ABCC5
HTR3D
HTR3D
HTR3D
DVL3
HTR3E
VPS8
VPS8
C3orf70
EHHADH
EHHADH
MIR5588
MAP3K13
MAP3K13
MAP3K13
MAP3K13
TRA2B
TRA2B
ETV5
DGKG
DGKG
DGKG
CRYGS
TBCCD1
TBCCD1
TBCCD1
AHSG
HRG
RFC4
RTP1
MASP1
MASP1
RTP4
SST
RTP2
BCL6
BCL6
BCL6

Built-in genome browser session



You can use your own Hi-C data
- no need to upload!



Welcome to the Interactive Hi-C Data Browser!

Accessing Hi-C Data

Step 1. Choose the assembly of Hi-C data:

Species: human
Assembly: hg19

Step 2. Choose the source of the data:

● Browse Available Hi-C Data

Tissue: GM12878
Resolution: 25kb

OR

○ Use Your Own Data

Data URL

Step 3. Choose the region to show Hi-C interactions:

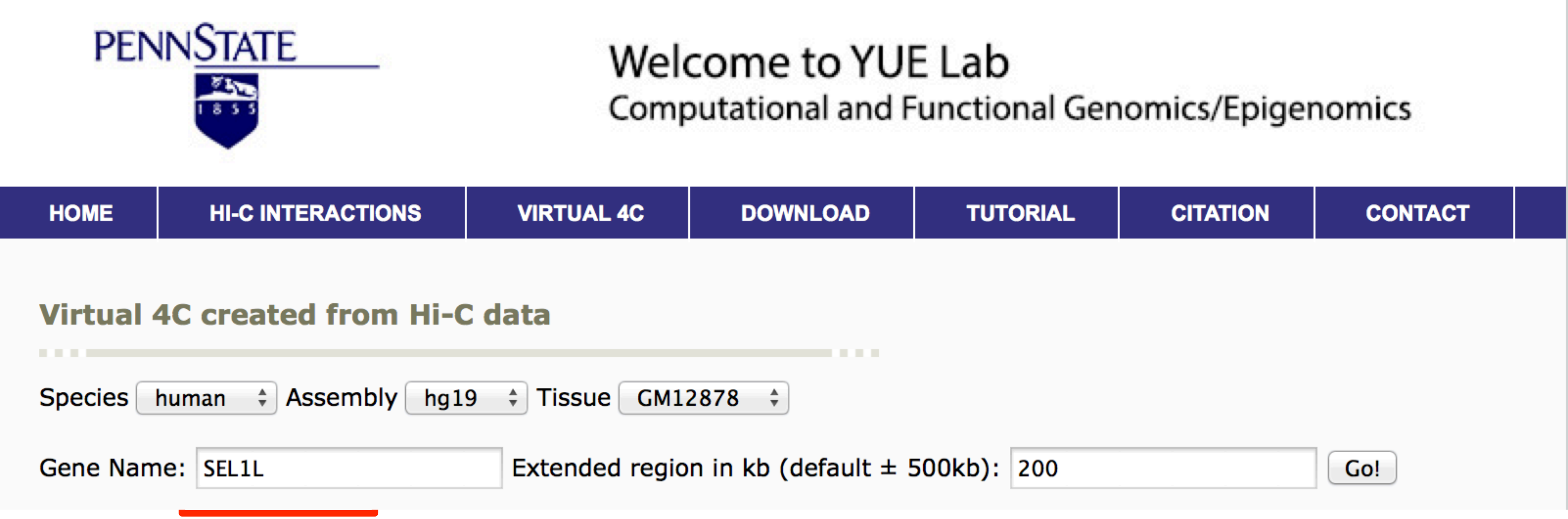
Option 1: Search by Gene Name	Option 2: Search by Location
Gene/Genomic Feature	Chromosome: chr1 Start End

Optional: User-defined UCSC Browser Session

UCSC Genome Browser Session ID

To retain your UCSC custom tracks and settings, save this URL and paste it here for the next time.
We recommend creating a session at UCSC and paste the URL instead. (Guide: http://genome.ucsc.edu/goldenPath/help/hgSessionHelp.html)

Use virtual 4C to hypothesize the enhancer promoter interaction



Welcome to YUE Lab

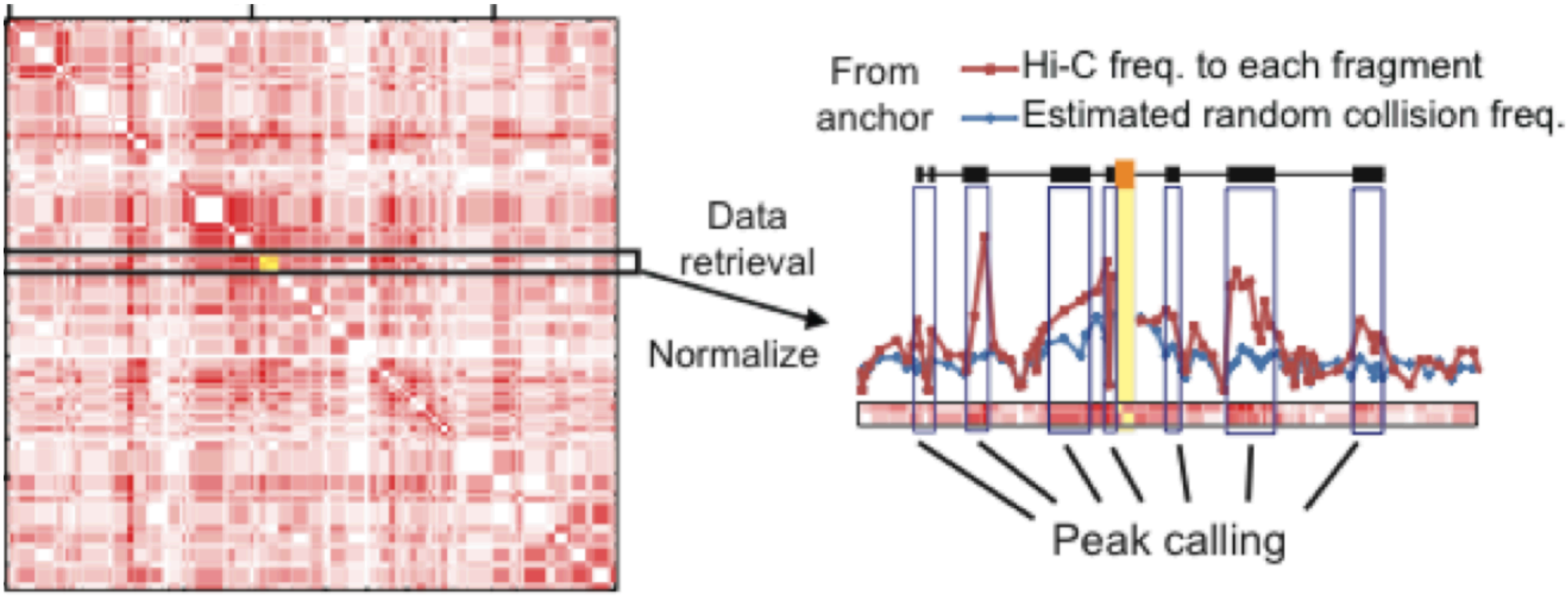
Computational and Functional Genomics/Epigenomics

Virtual 4C created from Hi-C data

Species: human
Assembly: hg19
Tissue: GM12878

Gene Name: SEL1L
Extended region in kb (default ± 500kb): 200

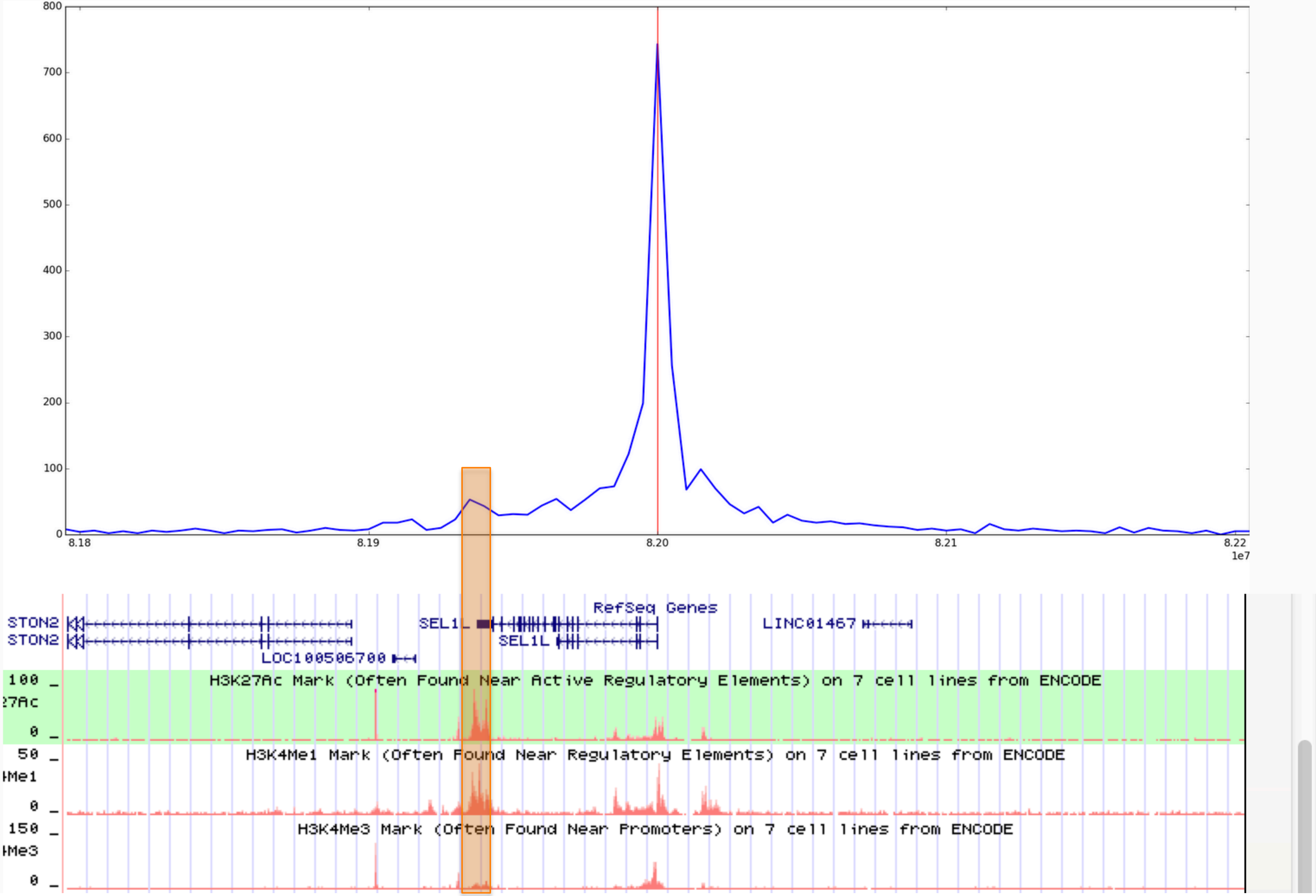
Virtual 4C is derived from Hi-C matrix



Jin et al, Nature, 2013

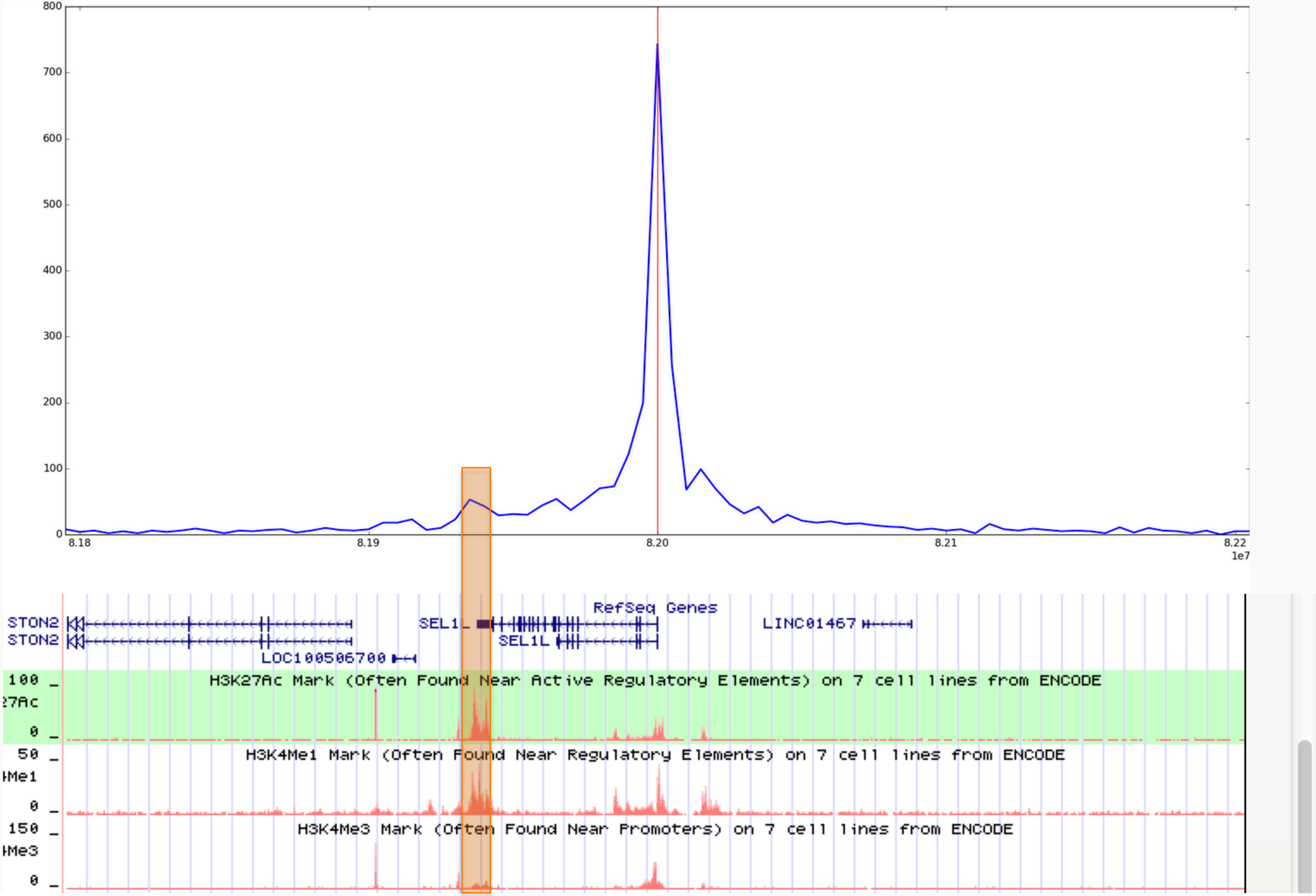
Available smallest resolution for hg19 GM12878 is 5 kb from Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014;159:1665-1680.

Gene **SEL1L** (TSA305, UNQ128/PRO1063) [ENSG00000071537] with TSS at chr14:82000205.

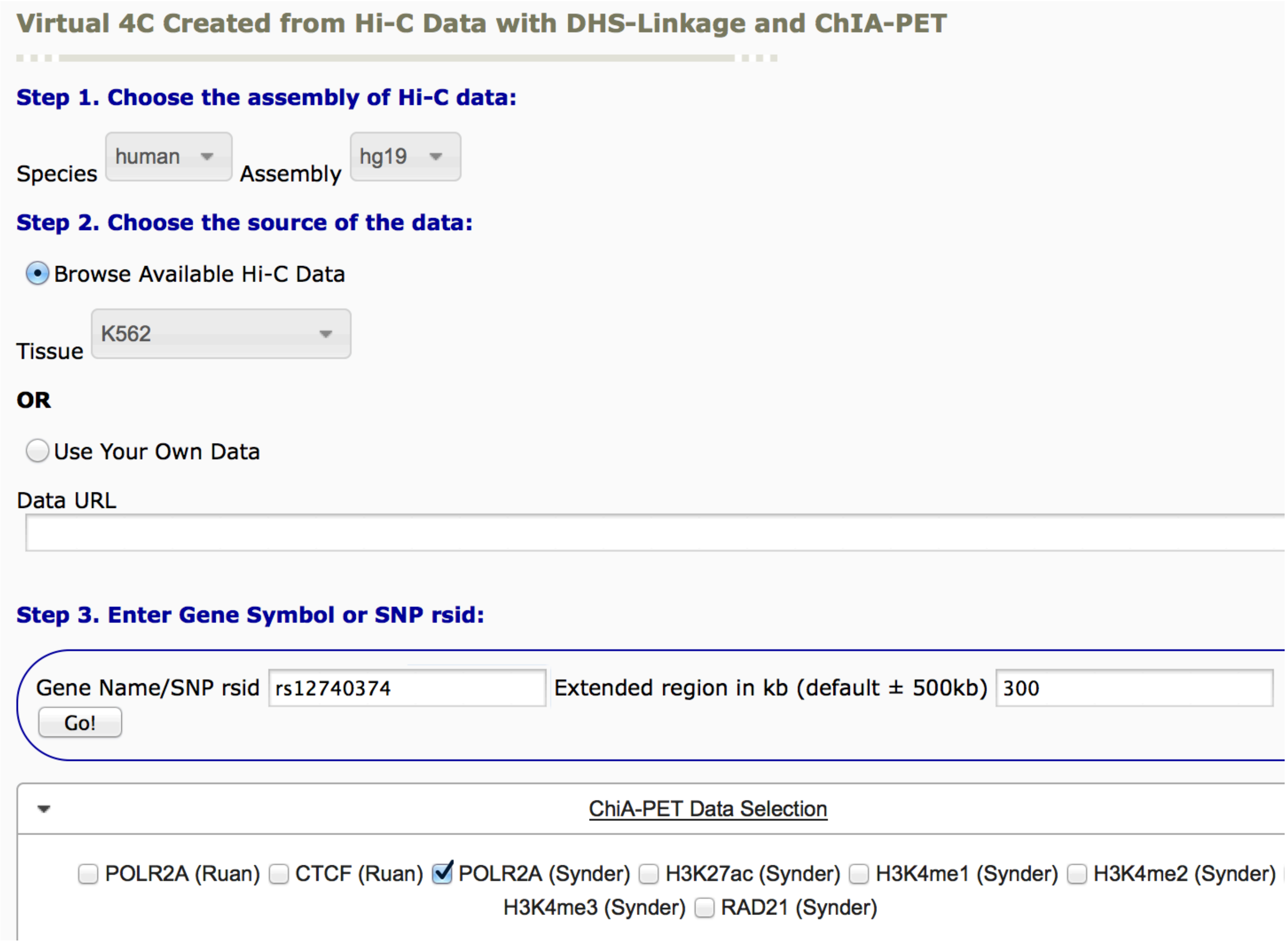


Available smallest resolution for hg19 GM12878 is 5 kb from Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014;159:1665-1680.

Gene **SEL1L** (TSA305, UNQ128/PRO1063) [ENSG00000071537] with TSS at chr14:82000205.



Define target genes for rs12740374



Virtual 4C Created from Hi-C Data with DHS-Linkage and ChIA-PET

Step 1. Choose the assembly of Hi-C data:

Species: human

Assembly: hg19

Step 2. Choose the source of the data:

● Browse Available Hi-C Data

Tissue: K562

OR

○ Use Your Own Data

Data URL:

Step 3. Enter Gene Symbol or SNP rsid:

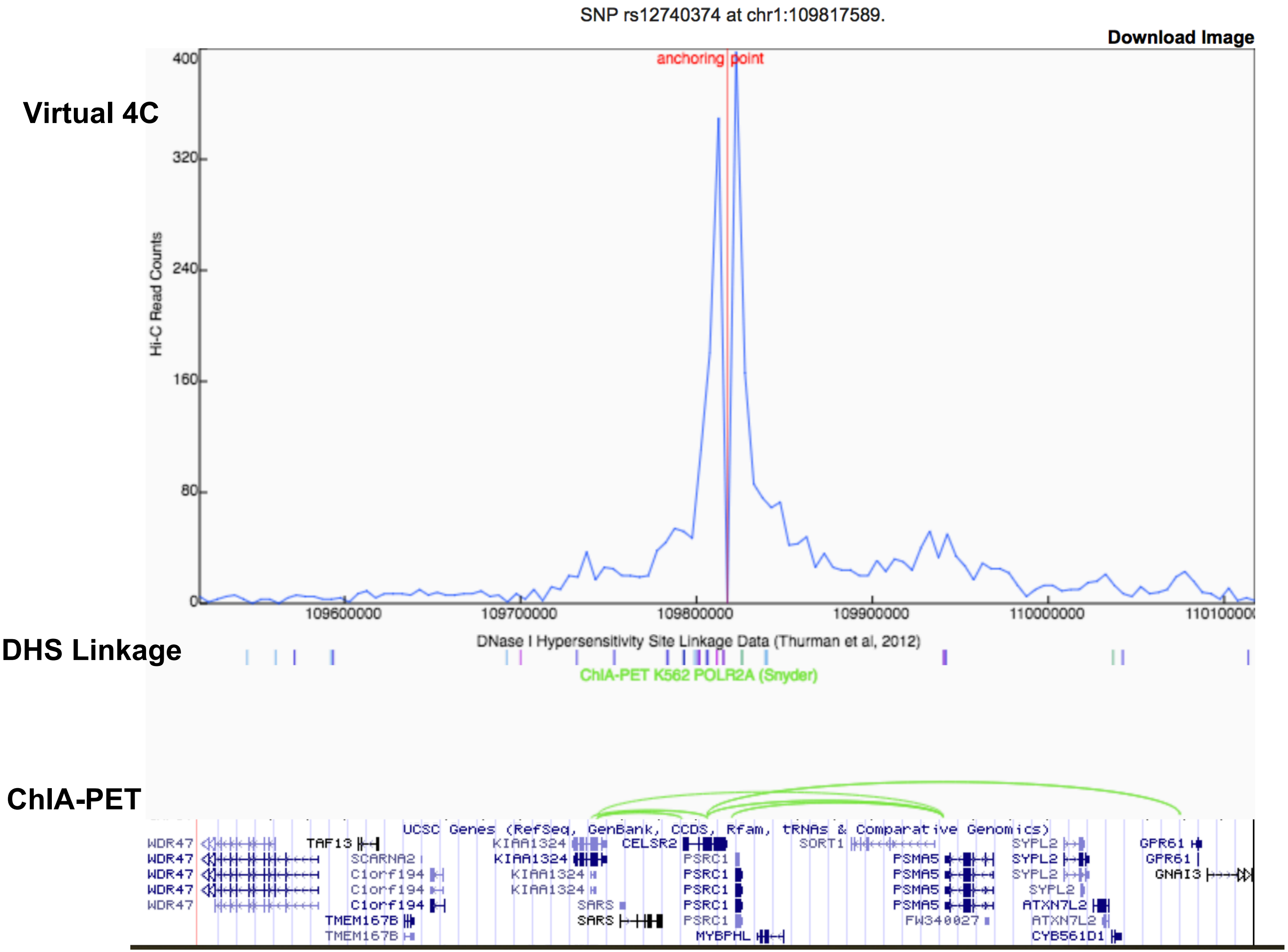
Gene Name/SNP rsid: rs12740374

Extended region in kb (default ± 500kb): 300

ChiA-PET Data Selection

☐ POLR2A (Ruan) ☐ CTCF (Ruan) ☑ POLR2A (Synder) ☐ H3K27ac (Synder) ☐ H3K4me1 (Synder) ☐ H3K4me2 (Synder)
H3K4me3 (Synder) ☐ RAD21 (Synder)

Define target genes for rs12740374



Summary

1. Brief tour of ENCODE project;
2. Search and download through ENCODE portal
3. Search and visualize ENCODE annotations (gene expression, promoters, enhancers …)
4. Visualize ENCODE chromatin interaction data (Hi-C, ChIA-PET)

Acknowledgement

ENCODE consortium

4D Nucleome consortium

ENCODE DCC

- Mike Cherry

…

ENCODE EDAC

- Zhiping Weng

…

Dr. Bing Ren

Yue Lab

- Yanli Wang
- Lijun Zhang
- GalYaroslavsky
- Tyler Derr

and many others …